# Measurement Error And Causal Inference With Instrumental Variables

**Manoochehr Babanezhad**

**Supervisor: Prof. dr. Els Goetghebeur**

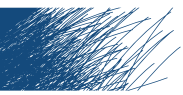
**Supervisor: Prof. dr. Stijn Vansteelandt**

In Partial Fulfillment of the Requirements for the Degree
**Doctor of Mathematics**

Academic year 2008-2009

Faculty of Sciences

Department of Applied Mathematics and Computer Science




# Contents

# Introduction

Causal inference deals with cause-effect relationships between interventions (exposures) and outcomes (responses) in many fields of study. It aims to infer causal effects of intervention (exposure) from empirical data on outcomes of interest through design and analysis. For example in the medical sciences, one might be interested to estimate the causal effect of a specific drug or other medical intervention on a primary health or risk outcome. In nutritional studies, one may investigate the effects of nutrition habits on cancer or chronic illness. In bioassay experiments, one wishes to assess the effects of herbicides on the dry weight of plants grown in the same pot. In social epidemiology, one may attempt to understand the relationship of socio-economic status with health, and ponder to find the effect on all-cause mortality. For this purpose, large amounts of data are required to sample through experimental or observational studies and statistical techniques are used to analyze these data to infer the causal effects of interest.

For doing so, investigators might be confronted with the difficulty of determining valid measurements of exposure. Specifically, in some situations exposure is hard to measure and researchers may not succeed to observe the true exposure for each subject in practice. For instance, in the context of noncompliance adjustment in randomized controlled trials, simple measures of compliance with drug therapy, such as pill counts are notorious for overestimating the amount of drug actually taken. In nutritional studies for example, it is rational to believe that exposures are not precisely measured if they are obtained through self-report or questionnaires or even by technical sophisticated tools. In environmental problems individual levels of pollution and radiation are difficult to measure. In herbicide studies the amount of the herbicide actually absorbed by the plant is a quantity which cannot be accurately measured. Systematic measurement errors on exposure thus occur frequently and are inevitable in practice. Measurement error in exposure forms a common source of bias in exposure effect estimates on outcome. When no adjustments are made for measurement errors, the bias of effect estimates can grow unexpectedly large and lead to a loss of efficiency,

in particular when the exposure effect is confounded by measured or unmeasured covariates. Briefly, confounding of the exposure effect occurs when the outcome is influenced by prognostic factors of the exposure other than the exposure itself. These factors are called confounders.
New concerns over the impact of measurement error have arisen in the context of recently developed causal models for the noncompliance adjustment (Dunn, 1999; Goetghebeur and Vansteelandt, 2005). Investigating the impact of systematic measurement error in exposure or error of exposure misclassification in causal models and correcting for it, forms the topic of this thesis. To the best of our knowledge, this phenomenon has not been addressed yet in the causal inference literature.
Over the past decades much attention has been paid to the impact of measurement error and of misclassification error in many of the well known regression models (e.g., Fuller, 1987; Gustafson, 2003; Carroll et al., 2006). In particular, estimation strategies have been proposed that successfully correct estimated exposure effects for error-prone exposure measurements. In chapter 1, we give an introduction to regression modeling in settings where some explanatory variables are measured with error. We explain the impact of measurement error in continuous and categorical explanatory variables. Specifically, we illustrate for instance that, in simple linear regression of outcome on explanatory variable, an error-prone explanatory variable leads not only to less precise estimates of regression coefficients but it also biases them towards zero. Moreover, we introduce various methods of adjustment for measurement error that can be used if its magnitude may be estimated from supplementary information.
In chapter 2, we will give an introduction to the statistical framework of counterfactual outcomes for causal inference as it has been developed over the past decades. We define causal effects and causal models formally in this setting, and state the assumptions which allow to identify causal effects. We explain the problem of confounding and discuss methods to control for measured as well as unmeasured confounding. We introduce the special clan of problems of noncompliance in randomized controlled trials and expand on methods to adjust for it. We describe the instrumental variables (IVs) approach to allow for inference on the causal effect of exposure on outcome in the presence of noncompliance in randomized controlled trials, and for unmeasured confounder adjustment in observational studies. We continue to investigate the problem of measurement error in exposure and important practical implications under linear structural mean models (Robins, 1994; Goetghebeur and Lapp, 1997).
In chapter 3, we elaborate on instrumental variable (IV) estimators for the causal effect of an exposure when the outcome is dichotomous. Specifically, we give an expository review of exact as well as approximate IV-estimators for the causal odds ratio, that have been proposed in the biostatistical, epidemiological and econometric literature. Methods comparisons are made, both theoretically and via extensive simu-

lation, and new insights are developed into the assumptions underlying their validity. The different estimators are used to assess the risk of gastrointestinal (GI) complications attributable to different non-steroidal anti-inflammatory drugs (instead of Cox-2 inhibitors).
In chapter 4, we will calculate biases and develop analytic methods to correct estimators for systematic measurement error in a continuous exposure under linear structural mean models for unconstrained outcomes. We focus on the impact of systematic error in compliance measurements on compliance adjusted analyses. Specifically, we build on ideas from linear regression models with error in the covariates to show how an IV for the measurement error can help correct IV-based causal effect estimators for systematic error under linear structural mean models.
In chapter 5, we explore the consequences of misclassification error on a dichotomous exposure under causal models, including alternatives to ordinary regression adjustment for confounder control. We focus on inverse probability of treatment weighted (IPTW) estimators for the parameters indexing marginal structural mean models and on G-estimators and propensity score adjusted estimators under semiparametric causal models when the exposure is subject to misclassification. We quantify the asymptotic bias of causal effect estimators in terms of misclassification probabilities depending on covariates. Furthermore, we formulate misclassification of a time-varying exposures at each time $t$ in longitudinal repeated measures data. In chapter 6, we will review the main results of this thesis, give a final discussion and plans for future work.
Chapters 3, 4 and 5 were originally written as stand-alone articles. Chapter 3 has been submitted to Statistical Science (Babanezhad, Vansteelandt and Goetghebeur, 2009). Chapter 4 has been accepted for publication in Statistica Sinica (Vansteelandt, Babanezhad and Goetghebeur, 2008). Chapter 5 has been submitted to the Journal of Statistical Planning and Inference (Babanezhad, Vansteelandt and Goetghebeur, 2008). The results have been presented at international conferences. The notations are introduced per chapter and may therefore differ throughout the complete thesis. While the statistical discussion and development throughout this thesis is general, our examples are drawn from biostatistics and epidemiology.

# Chapter 1

# Covariate Measurement Error in Regression Models


## Summary

This chapter gives an introduction to regression modeling in settings where some explanatory variables are measured with error. The purpose is to provide a summary of measurement error problems in explanatory variables and of analysis strategies for correcting this, focusing on traditional regression models. We start by illustrating how measurement error or misclassification in explanatory variables occurs in real situations. We then proceed with basic definitions of error structure in continuous and categorical explanatory variables and with measurement error assumptions. Specifically, we investigate the impact of measurement error and, in particular, misclassification on parameter estimates in linear regression analysis. We introduce various methods of adjustment for measurement error that can be used if its magnitude may be estimated from supplementary information.


## 1 Introduction and problem setting

Regression analysis is a statistical methodology that utilizes the relationship between two or more variables so that an outcome (response) variable can be predicted from the explanatory variables (predictors). The so-called 'measurement error problem' or 'errors-in-variables problem' considered in this chapter arises in situations

where the explanatory variable $X$ is difficult to measure or cannot be accurately measured for all study subjects. In particular, random or systematic errors occur in measurements of explanatory variables in a variety of research fields. Systematic measurement error is referred as a persistent error having zero or nonzero mean that can be attributed to inaccuracy inherent in the systems of measurement. This error may occur at various stages of the data collection. It may occur due to imperfect methods of measuring, or the result of misreporting by subjects, or miscoding by the collectors of the data, or incorrect transformation from initial reports into a form ready for analysis, and so on. One should distinguish the systematic error in the measurements of explanatory variable $X$ from the random error term in regression models, which is commonly added to the assumed relationship between outcome $Y$ and explanatory variable $X$ to capture the influence of everything else on $Y$ other than $X$. In particular, the effect of the random model errors will reduce with increasing sample size. In contrast, the implications of systematic measurement error in the explanatory variables does not reduce with increasing sample size (Fuller, 1987; Fosgate, 2006).

Throughout this chapter, we refer to error in a continuous explanatory variable $X$ as plain measurement error. When $X$ is discrete (categorical), which is often the case in medical, epidemiological and biostatistical applications, then misclassification is the term to use. This is generally studied separately from measurement error in continuous explanatory variables, although there is clearly much overlap. In both cases, $X$ is also called the 'error-prone' or 'mismeasured variable'. The following examples illustrate how measurement error or misclassification in explanatory variables occurs in real-life situations.

**Example 1**. The effect of nutrition habits on cancer, such as breast and colon cancer has been well investigated in the literature (e.g., Kipnis et al., 2003; Carroll et al., 2006). Because it is both difficult and expensive to measure long-term diet in a large cohort, instead of observing long-term diet, researchers typically measure a 24-hours recall. That is, each subject's diet in the previous 24-hours was recalled and nutrition variables were computed on the basis of this. Measurement error in nutrient instruments can be very large, for example because of the daily and seasonal variability of an individual's diet.

**Example 2**. Alzheimer's disease is the most common form of dementia. It is thought that the level of aluminium deposits, which may build up in the brain over time, have an effect on an evaluation score for diagnosing an individual developing Alzheimer's (Campbell, 2002; Thompson and Carter, 2007). If investigators wanted to estimate the association between the risk of Alzheimer's disease and the level of aluminium deposits in the brain, problems would arise because a perfectly accurate measure of aluminium levels is simply not attainable.

**Example 3**. Bashir et al. (1997) discuss a study examining the relationship between the incidence of minor Ischaemic Stroke and levels of haemostatic factors which retard the flow of blood in blood vessels. The study follows a case-control design, comparing a group of subjects who had a minor stroke to a group that did not. Initial blood samples were assayed twice, giving two measurements of each of the explanatory variables. Also, one-year follow-up blood samples were obtained for some of the control subjects, and these were also assayed twice. As is typical for biochemical variables, mismeasurement is at play in two ways. First, there is pure laboratory error which leads to two different numerical measurements for the same blood sample. Second, levels of the haemostatic factors vary somewhat from day-to-day within a given subject. Operationally, it makes sense to define the explanatory variable as the subject's average level, but measurement error arises because of the day-to-day fluctuations (Gustafson, 2003). Even in the absence of pure laboratory error then, two blood samples taken on different days are not likely to give identical measurements.
**Example 4**. The Framingham Heart study is a large cohort study, which follows individuals for the development of coronary heart disease. The explanatory variables are systolic blood pressure (SBP) and serum cholesterol, both of which are subject to measurement error (e.g., MacMahon et al., 1990; Carroll et al., 2006). In particular, it is impossible to measure long-term systolic blood pressure because blood pressure measurements are well known to have major daily as well as seasonal variation.

It follows from the above examples that, researchers may not succeed to observe the true explanatory variable $X_i$ for each subject study $i$ in practice. They instead observe a variable $W_i$ which approximates, but may differ from $X_i$ for each subject study $i$. The observed explanatory variable $W_i$ is related to $X_i$ in terms of an error model, as described in the next section. The goal of measurement error modeling is to obtain nearly unbiased estimates of explanatory variable effects and valid inferences on the outcome of interest. Attainment of this goal requires careful analysis. While it is tempting to simply plug-in $W_i$ instead of $X_i$, making no further adjustment in the usual fitting methods typically leads to biased estimates and misleading statistical inferences (Schneeweiss and Mittag 1986; Fuller, 1987; Carroll and Stefanski, 1990; Stefanski and Buzas 1995; Gustafson, 2003; Dunn, 2005; Buzas et al., 2005; Carroll et al., 2006). In view of this, in this chapter, we shed further light on the known standard approach for correcting the impact of measurement error in regression models. Although outcome variables may themselves be subject to measurement error, our attention is limited to measurement error in explanatory variables as the impact of the latter type of error is typically more severe.

# 2 Models for measurement error

Determining the structure of measurement error or misclassification in explanatory variables under regression models initially requires specifying models for the error process. Such models quantify the relationship between the true explanatory variable $X$ and the observed explanatory variable $W$. In particular, they describe the structure of systematic error in measurements of continuous or discrete (categorical) explanatory variables. In the literature on the measurement error problem, two general types of measurement error model are commonly considered for continuous $X$, where the error in the explanatory variable assessment is the difference between the observed explanatory variable and the true explanatory variable. For discrete $X$, the measurement error model can be defined in terms of conditional misclassification probabilities.

## 2.1 Classical error model

In cases where explicit attention is paid to measurement error, the standard error model is typically the 'Classical measurement error model' or 'Classical additive measurement error model'. This is well suited to describe the situation where the true explanatory variable $X$ is imperfectly recorded by $W$,

$$W = X + U$$

where $U$, the measurement error, is assumed to be independent of $X$; that is, $U \perp\!\!\!\perp X$. The classical measurement error model in its simplest form is appropriate when an attempt is made to determine $X$ directly. Errors of the classical type arise when a quantity is measured by some device and repeated measurements vary around the true value. For example, consider the measurement of systolic blood pressure in example 4, which is known to have daily and seasonal variations. In trying to measure long-term systolic blood pressure, the true long-term blood pressure can be considered fixed for an individual; the measured value is then perturbed by error. In cases like this one, it makes sense to use the classical error model (Carroll et al., 2006). In practice, it is common to assume that the measurement error $U$ has mean 0 so that $\mathrm{E}(U|X) = \mathrm{E}(U) = 0$. This implies that $\mathrm{E}(W|X) = X$, suggesting that $W$ is an unbiased measure of $X$. In fact, virtually all developments on measurement error assume the error to be normally distributed with mean zero and of the classical type (i.e., independent of the explanatory variable $X$). Moreover, the error structure of $U$ could be homoscedastic (constant variance) or heteroscedastic. Note that not all measurement methods assume unbiased measurements. For instance, a slightly more

general error model that allows for systematic error is (Freedman et al., 2008),

$$W = \gamma_0 + \gamma_1 X + U,$$

where $U$ is independent of $X$: with $\mathrm{E}(U) = \mathrm{E}(U|X) = 0$. This implies that $\mathrm{E}(W|X) = \gamma_0 + \gamma_1 X$, unlike the classical error model, suggesting that $W$ is a possibly biased measure of $X$. This model is motivated by dietary self-report data that appear to conform to this model after a suitable transformation (Kipnis et al., 2003). This error model is referred to as the 'Non-classical measurement error model' or 'Error calibration model' to distinguish it from the classical measurement error model where $\gamma_0 = 0$ and $\gamma_1 = 1$. The term calibration means that $W$ has to be calibrated to make it unbiased for $X$, for instance, by using $(W - \gamma_0)/\gamma_1$.

## 2.2 Berkson error model

The 'Berkson error model' or the 'Controlled variable model' is an alternative to the classical measurement error model, where $X$ varies around $W$. It is based on the assumption that the measurement error is independent of the observed explanatory variable $W$, in the sense that

$$X = W + U$$

where $U$ is independent of $W$; that is, $U \perp\!\!\!\perp W$. The Berkson error model has been found to be useful in agricultural and medical studies. As an example, consider the herbicide study (Rudemo et al., 1989; Koul and Song 2008) in which a nominal measured amount $W$ of herbicide was applied to a plant but the actual amount absorbed by the plant $X$ is unobservable. Here, the actual amount absorbed by the plant varies around the nominal measured amount $W$ due to error. Other examples have been studied where the relations between the yield of a crop or the efficacy of a drug, $Y$, and the amount of a fertilizer or drug used, $X$, (Wang, 2004; Carroll et al., 2006). Suppose the fertilizer or the drug is applied at predetermined doses $W$. The actual absorption of the fertilizer in the crop or the drug in the patient's blood may vary randomly around the set doses, because of the local earth conditions or the individual biological conditions. In these cases, if the amount of $W$ is properly calibrated, then the actual absorption $X$ will vary randomly around $W$.
In practice, it is common to assume that the measurement error $U$ has mean 0 so that $\mathrm{E}(X|W) = W$, suggesting that $X$ is unbiased for $W$. Deviation from this can be allowed, for instance by using the error calibration model

$$X = \gamma_0 + \gamma_1 W + U,$$

where $\mathrm{E}(U) = \mathrm{E}(U|W) = 0$. This model encompasses the Berkson error model, which corresponds to $\gamma_0 = 0$ and $\gamma_1 = 1$.
Many statistical methods in the literature on measurement error modeling typically suppose that the error follows the classical measurement error model. However, determining an appropriate error model to use in the data analysis depends upon the circumstance and the available data.

## 2.3 Error of misclassification

Many explanatory variables encountered in statistical practice are discrete (categorical), rather than continuous (e.g., Spiegelman et al., 1995; Gustafson, 2003; Lederer and Küchenhoff, 2006). In many epidemiologic applications, such categorical exposures (e.g., level of smoking, dietary in take, quintiles of fat, etc.), for instance those obtained through self-report or questionnaires, may be error-prone. Misclassification may also arise by transforming a continuous variable into a discrete explanatory variable. Misclassification error basically differs from measurement error as discussed in previous section because the observed explanatory variable $W$ cannot be expressed as a sum of the true explanatory variable $X$ with an error variable. Rather, one must characterize the measurement error in terms of misclassification probabilities.
**Example 5**. Consider an example where $X$ and $Y$ denote the presence of a particular mental disorder (MD) such as major depression in parents and in offspring, respectively in prospective study (Höfler, 2005). The magnitude of the association between $X$ and $Y$ here would represent the degree of familial aggregation of the disorder under consideration. Suppose that $W$ is obtained using an error-prone method of deriving a diagnosis for MD in the parents. Then there are two misclassification probabilities in each variable: $P(W = 1|X = 0)$ denotes the conditional probability that MD is observed in the parents when in fact there is no MD (false positive rate); and likewise, $P(W = 0|X = 1)$ is the probability that there is apparently no MD, while the diagnostic criteria are actually met (false negative rate).
The likely extent of misclassification of categorical variables is usually specified in terms of conditional probabilities of misclassification. These conditional probabilities are often expressed as the probability of the observed explanatory variable $W$ given the true explanatory variable $X$; that is, $P(W|X)$. For dichotomous variables, it is conventional to express these through the sensitivity, $\pi_{1|1} = P(W = 1|X = 1)$, and the specificity, $\pi_{0|0} = P(W = 0|X = 0)$, of the classification. In case-control studies, the sensitivity refers to the probability of correctly classifying a truly exposed subject as exposed and the specificity to the probability of correctly classifying an unexposed subject as unexposed. Thus if sensitivity, for instance, is 0.8 and specificity is 0.7, the probability of misclassifying an exposed subject as unexposed is $1 - 0.8 = 0.2$, and

the probability of misclassifying an unexposed subject as exposed is $1 - 0.7 = 0.3$. For categorical variables of more than two levels, many different sorts of misclassification can occur, which can be specified in a matrix of misclassification probabilities (Morrissey and Spiegelman 1999; Küchenhoff et al., 2006).

# 3 Regression models and the impact of measurement error

Measurement error analysis generally involves three parts. Part 1 consists of a measurement error model describing the relationship between the observed explanatory variable $W$ and the unobservable true explanatory variable $X$. Part 2 consists of the regression model of interest for the association of outcome $Y$ with $X$, possibly adjusted for covariates $Z$ which are measured without error. Part 3 is linking $Y$ and $W$, possibly adjusted for covariates $Z$, under measurement error assumptions such as nondifferential measurement error (explained in the next section). However, to correct for bias due to measurement error or misclassification, some further assumptions are typically required regarding the distribution of the unobservable true explanatory variable $X$ and the error term $U$. Most methods for measurement error analysis have been worked out for ordinary linear regression (Fuller, 1987) and common nonlinear regression models, such as the logistic regression model (Gustafson, 2003; Carroll et al., 2006). In this section, we develop a general regression model formulation, but focus eventually on the simple regression model as an illustrative example. In particular, let us consider the following general regression model for the association of outcome $Y_i$ with a mismeasured explanatory variable $X_i$,

$$\mathrm{E}(Y_i|X_i, Z_i) = h(X_i, Z_i; \beta^*), \tag{1.1}$$

where covariate $Z_i$ is measured without error, $h(X_i, Z_i; \beta)$ is a known function smooth in $\beta$, and $\beta^*$ is an unknown finite-dimensional parameter. Given explanatory variable $X_i$, $h$ could be the identity function when the outcome $Y_i$ is continuous or the logistic function when $Y_i$ is dichotomous.

## 3.1 Structural and functional models

In the measurement error analysis, it is typically necessary to make assumptions about the distribution of $X$. Specifically, one may consider the $X$'s to be unknown, nonrandom constants or random variables with a distribution given by a density function $f_X(x; \gamma^*)$, where $\gamma^*$ may be a vector of nuisance parameters describing the

distribution of $X$. In so-called structural models, the explanatory variable $X_i$ is regarded as a random variable; thus $X_i$ is assumed to be independent random draws from a distribution for each $i = 1, ..., n$. In contrast, in so-called functional models, $X_i$ is regarded as unknown fixed constant for each $i = 1, ..., n$. Throughout, we mainly devote attention to the structural model, but it is nonetheless important to say that functional modeling is attractive because it avoids assumptions.

### 3.2 Differential and nondifferential measurement error

Measurement error can be either differential or nondifferential. Nondifferential measurement error in the presence of an error-free covariate $Z$ occurs when $Y$ is independent of $W$, given $X$ and $Z$; that is, $Y \perp\!\!\!\perp W|X, Z$ so that $f_{Y|W,X,Z} = f_{Y|X,Z}$. The observed measurement $W$ is then said to be a surrogate for $X$. If this assumption fails, then the error is said to be differential. The assumption of nondifferential measurement error thus states that $W$ contains no information for predicting $Y$ in addition to the information already contained in $X$ and $Z$. Intuitively, it thus suggests that the measurement error arises in a manner which is blind to the outcome variable, so in some sense the problem is limited to one of a difficulty in making measurements. This assumption is useful (Buzas et al., 2005; Greenland and Gustafson 2006; Carroll et al., 2006), because it greatly simplifies the link between the association of $Y$ and $W$ and the association of $Y$ and $X$ (see Section 4). Many statistical methods in the literature on measurement error modeling are therefore based on the assumption of nondifferential measurement error. However, it is important to understand this concept and to recognize when it is an appropriate assumption and when it is not. Nondifferential error is plausible in many cases. In example 4 of Section 1, $X$ refers to long-term systolic blood pressure, whereas $W$ denotes the blood pressure measurement on a single day. It is reasonable to believe that a single day's blood pressure contains no more information than long-term blood pressure (Carroll et al., 2006), and hence that measurement error is nondifferential. Throughout, we focus on nondifferential measurement error unless stated otherwise.

## 4 Continuous explanatory variables

We begin with an illustration of the effects of measurement error for the case of homoscedastic ordinary linear regression where the explanatory variable $X$ is continuous. Consider the multiple linear regression model,

$$\mathrm{E}(Y_i|X_i, Z_i) = \beta_0^* + \beta_1^* X_i + \beta_2^* Z_i, \tag{1.2}$$

where $Z_i$ is an error-free covariate and $X_i$ is an error-prone explanatory variable with mean $\mu_x$ and variance $\sigma_x^2$. Here, $\beta^* = (\beta_0^*, \beta_1^*, \beta^*)$ is an unknown finite-dimensional parameter, with $\beta_1^*$ encoding the conditional association between $X_i$ and $Y_i$ (given $Z_i$). Under the classical additive measurement error model, $W_i = X_i + U_i$, we observe a variable $W_i$ instead of $X_i$, where the measurement error $U_i$ is independent of $(X_i, Z_i)$ with mean zero $\mu_u = 0$. In this section, we will additionally assume $U_i$ to be normally distributed with constant variance $\sigma_u^2$ which is assumed known or can be estimated from supplementary data (next section). Suppose that the primary interest of the study lies in the conditional association $\beta_1^*$ of $X$ and $Y$. When the investigator is unaware of the measurement error or chooses to ignore it, he/she may simply regress $Y$ on $(W, Z)$, and would then not obtain a consistent estimate of $\beta^*$, but instead obtain an estimate of $\theta^* = (\theta_0^*, \theta_1^*, \theta_2^*)$ indexing the following regression model,

$$\mathrm{E}(Y_i|W_i, Z_i) = \theta_0^* + \theta_1^* W_i + \theta_2^* Z_i. \tag{1.3}$$

Throughout, model (1.3) is called the naive model and parameter $\theta^*$ is called the naive parameter. The latter is implied by model (1.2) by the fact that the relationship between $Y$ and $(W, Z)$ is greatly simplified when the measurement error is nondifferential:

$$\begin{aligned} \mathrm{E}(Y|W, Z) &= \mathrm{E}\{\mathrm{E}(Y|W, Z, X)|W, Z\} \\ &= \mathrm{E}\{\mathrm{E}(Y|X, Z)|W, Z\} = \beta_0^* + \beta_1^* \mathrm{E}(X|W, Z) + \beta_2^* Z. \end{aligned} \tag{1.4}$$

The latter implies that the regression of $Y$ on $(W, Z)$ is equal to the regression of $Y$ on $\{\mathrm{E}(X|W, Z), Z\}$. In statistics, a great deal of current research in parametric and semiparametric statistical inference is organized around estimating equations. In regression models, one often solves the estimating equation corresponds to that regression model to obtain regression coefficient estimates. The estimating equation is then called unbiased if it has expectation zero when evaluated at the true parameter values. If it is unbiased, its solution is a consistent and asymptotically normal estimator for the considered parameters. In the absence of measurement error a consistent and asymptotically normal estimator for $\beta^*$ under model (1.2) can be obtained by solving,

$$0 = \sum_{i=1}^{n} U_i(Y_i, X_i, Z_i; \beta^*) = \sum_{i=1}^{n} \begin{pmatrix} 1 \\ X_i \\ Z_i \end{pmatrix} (Y_i - \beta_0^* - \beta_1^* X_i - \beta_2^* Z_i)\,.$$

Here, $U_i(Y_i, X_i, Z_i; \beta^*)$ is an unbiased estimating equation, because $\mathrm{E}\{U_i(Y_i, X_i, Z_i; \beta^*)\} = 0$. By replacing $W$ instead of $X$, $U(Y_i, W_i, Z_i; \beta^*)$ may not generally be unbiased.

Throughout, we will use estimating equations to find the bias formula for the true coefficients due to measurement error. Under the naive model (1.3), the limiting parameter $\theta^*$ is obtained by solving the following expected estimating equation

$$0 = \mathrm{E}\{U_i(\theta^*)\} = \mathrm{E}\left\{ \begin{pmatrix} 1 \\ W_i \\ Z_i \end{pmatrix} (Y_i - \theta_0^* - \theta_1^* W_i - \theta_2^* Z_i) \right\}. \tag{1.5}$$

Comparing this with the true model (1.2) then yields a bias formula. Note that, throughout we use $U(\theta^*)$ to denote an estimating function, and $U$ for random measurement error.

## 4.1 Bias with the Classical error model

For simplicity, we first start with models that no error-free covariate. Consider linear model (1.2) with no error-free covariate $Z$ and assume the classical error model holds. As stated, with the classical additive error model, measurement error $U$ is independent of $X$ with mean 0 and variance $\sigma_u^2$. It follows then from the classical additive error model that; $\mathrm{E}(W|X) = X$, $Var(W|X) = \sigma_u^2$, $\mathrm{E}(W) = \mu_x$, $Var(W) = \sigma_x^2 + \sigma_u^2$, and $Cov(W, X) = Cov(X + U, X) = \sigma_x^2$. Under regression model (1.3) with no covariate $Z$, the naive coefficient estimators can be obtained by solving,

$$\begin{aligned} 0 &= \mathrm{E}\{U(\theta^*)\} \\ &= \mathrm{E}\left\{ \begin{pmatrix} 1 \\ W \end{pmatrix} (Y - \theta_0^* - \theta_1^* W) \right\} \\ &= \mathrm{E}\left[ \mathrm{E}\left\{ \begin{pmatrix} 1 \\ W \end{pmatrix} (Y - \theta_0^* - \theta_1^* W) \,|X, W \right\} \right]. \end{aligned}$$

This yields to the following equations under the nondifferential measurement error assumption,

$$\begin{cases} \mathrm{E}\{\mathrm{E}(Y|X) - \theta_0^* - \theta_1^* W\} = \mathrm{E}(\beta_0^* - \theta_0^* + \beta_1^* X - \theta_1^* W) = 0 \\ \\ \mathrm{E}[W\{\mathrm{E}(Y|X) - \theta_0^* - \theta_1^* W\}] = \mathrm{E}\{(\beta_0^* - \theta_0^*)W + \beta_1^* XW - \theta_1^* W^2\} = 0. \end{cases}$$

It follows from solving the above equations that;

$$\theta_1^* = \frac{Cov(W, X)}{Var(W)} \beta_1^* = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} \beta_1^*.$$

That is

$$\theta_1^* = \lambda \beta_1^* \tag{1.6}$$

where

$$\lambda = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} \tag{1.7}$$

and $\theta_0^* = \beta_0^* + (1-\lambda)\beta_1^* \mu_w$. As a result, ordinary least squares regression yields biased estimators of the regression slopes of error-prone explanatory variables. In particular, because $\lambda < 1$, the least squares regression coefficient $\theta_1^*$ is biased towards zero. This bias does not vanish with increasing sample size. In the measurement error literature, the attenuation factor $\lambda$ is called the 'reliability ratio' (variance of true explanatory variable divided by variance of measured explanatory variable, possibly given the error-free covariate); it expresses the degree of attenuation. It suggests that measurement error bias increases with decreasing explanatory variable variance. This usage of the term reliability is standard in technical discussions of measurement error, although it is used more generally among epidemiologists as a synonym for reproducibility or precision. If there is information on the magnitude of the error variance and the distribution of the explanatory variable $X$, then the above results allow in principle to correct for measurement error in estimating regression slopes, at least for reasonably simple forms of measurement error. An asymptotically unbiased estimator of $\beta_1^*$ is then given as follows,

$$\hat{\beta}_1 = \frac{\hat{\theta_1}}{\lambda} \tag{1.8}$$

where $\hat{\theta}_1$ is the ordinary least squares estimate of $\theta_1^*$. The resulting estimator (1.8) is sometimes called the regression coefficient corrected for attenuation. Further, $\mathrm{E}(\hat{\beta}_1) = \beta_1^*$ and $Var(\hat{\beta}_1) = Var(\hat{\theta}_1)/\lambda^2$. Because $\lambda < 1$, it is clear that $Var(\hat{\beta}_1) > Var(\hat{\theta}_1)$. This implies correcting for bias entails that the corrected estimator will be more variable than the biased estimator and then have wider confidence intervals. This generally means that the price for reduced bias is increased variance. This phenomenon is not restricted to the linear model; it occurs almost universally in the analysis of measurement error. Figure 1 displays the attenuation factor $\lambda$ as function of $\tau$ where $\tau^2 = \sigma_u^2/\sigma_x^2$. The first impression from this curve is that a moderate amount of measurement error does not cause a substantial attenuation (Gustafson, 2003; Carroll et al., 2006). For instance $\tau = 0.1$, interpreted as 10% measurement error yield $\lambda = 0.99$. Moreover, one would expect that because $W$ is not the true explanatory variable, it has a weaker relationship with the outcome than does $X$. This can be seen by the attenuation factor and also by the residual variance of regression of $Y$ on $W$. Indeed,

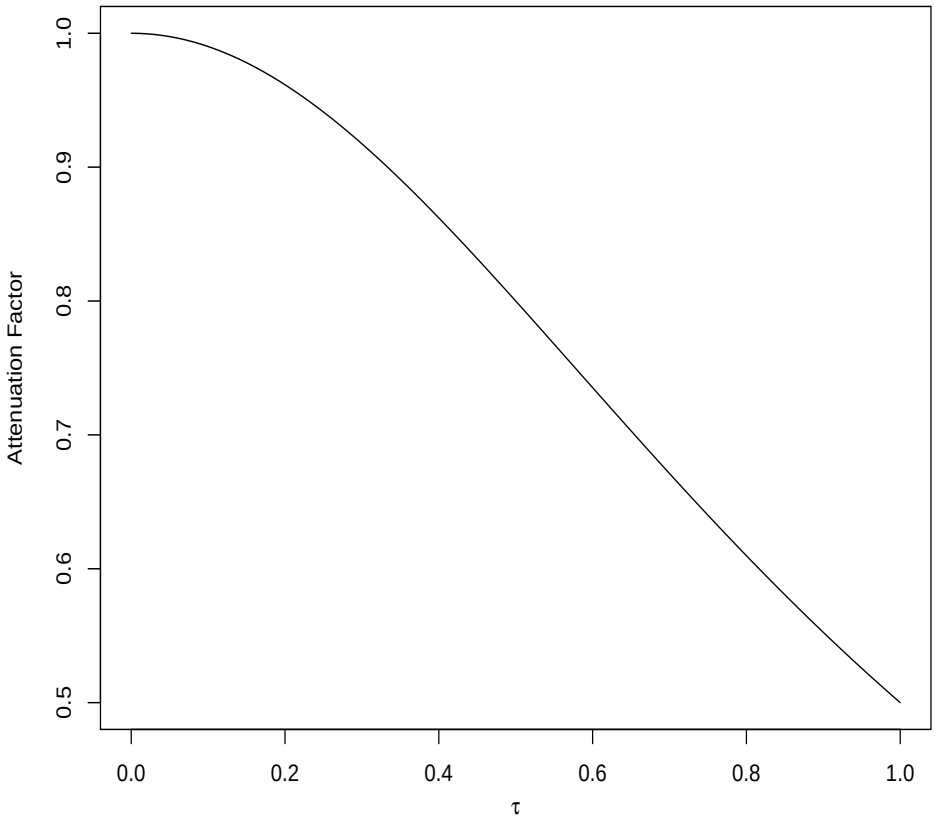



Figure 1: *Attenuation factor* $\lambda$ *as function of* $\tau$ *where* $\tau^2 = \sigma_u^2/\sigma_x^2$ *for linear regression.*

it follows from the normality assumption and the classical error model that

$$\begin{aligned} Var(Y|W) &= Var(\beta_0^* + \beta_1^* X + \epsilon|W) \\ &= \beta_1^{*2} Var(X|W) + \sigma_\epsilon^2 \\ &= \beta_1^{*2}(1-\rho_{xw}^2)\sigma_x^2 + \sigma_\epsilon^2 \\ &= \beta_1^{*2}(1-\frac{\sigma_x^2}{\sigma_w^2})\sigma_x^2 + \sigma_\epsilon^2 \\ &= \beta_1^{*2}\frac{\sigma_u^2\sigma_x^2}{\sigma_x^2+\sigma_u^2} + \sigma_\epsilon^2, \end{aligned}$$

where $\sigma_\epsilon^2$ is the variance of random error term in linear regression of $Y$ given $X$. We assume the normality assumption, $(X, U, \epsilon) \sim N\{(\mu_x, 0, 0)', \begin{pmatrix} \sigma_x^2 & 0 & 0 \\ 0 & \sigma_u^2 & 0 \\ 0 & 0 & \sigma_\epsilon^2 \end{pmatrix}\}$, that is, random vector $(X, U, \epsilon)$ is normally and independently distributed, where $\epsilon$ is the random error term in model (1.2). Consider now model (1.2) (with error-free covariate $Z$) with the classical error model. Under the nondifferential measurement error assumption

$$\begin{aligned} 0 = \mathrm{E}\{U(\theta^*)\} &= \mathrm{E}[\mathrm{E}\{U(\theta^*)|X, W, Z\}] \\ &= \mathrm{E}\left\{ \begin{pmatrix} 1 \\ W \\ Z \end{pmatrix} \mathrm{E}(Y|X,Z) - \theta_0^* - \theta_1^* W - \theta_2^* Z \right\}. \end{aligned}$$

As stated, with the classical error model measurement error $U$ is independent of $(X, Z)$ with mean 0 and variance $\sigma_u^2$. It then follows from the classical additive error model that $E(W|Z) = E(X|Z)$,

$$\mathrm{E}(XW|Z) = \mathrm{E}\{X(X+U)|Z\} = \mathrm{E}(X^2|Z)$$

and

$$Cov(W, X|Z) = Cov(X+U, X|Z) = \sigma_{x|z}^2.$$

Here we suppose that $X$ and $Z$ are linearly related

$$E(X|Z) = \eta_0^* + \eta_1^* Z.$$

In the Appendix, we show that by solving the latter expected estimating equation that,

$$\theta_1^* = \lambda_z \beta_1^*, \tag{1.9}$$

where

$$\lambda_z = \frac{\sigma^2_{x|z}}{\sigma^2_{x|z} + \sigma^2_u}, \tag{1.10}$$

and $\sigma^2_{x|z} = Var(X|Z)$. This implies that the bias induced in the regression coefficient for $X$ subject to measurement error is worsened by the presence of an additional precisely measured covariate $Z$, provided $X$ and $Z$ are correlated. The attenuation factor $\lambda_z$ is equal to $\lambda$ when $Z$ is independent of $X$. A similar derivation shows that the coefficient of $Z$ is also biased, unless $Z$ is independent of $X$. We obtain

$$\theta^*_2 = \beta^*_2 + (1 - \lambda_z)\beta^*_1\eta^*_1. \tag{1.11}$$

The bias of this regression slope is thus $(1 - \lambda_z)\beta^*_1\eta^*_1$. As a result, measurement error in $X$ can also induce bias in the regression coefficients of error-free covariates $Z$. This has important implications for analysis of covariance models in which the continuous explanatory variable is measured with error (Carroll, 1989; Buzas et al., 2005; Carroll et al., 2006). Note that hypothesis tests for the regression slope $\beta^*_1$ are valid in the presence of random, nondifferential measurement error (in the sense of preserving the nominal Type I error rate) because, there is no bias (see expression 1.9) under the null hypothesis that $\beta^*_1 = 0$. Tests may however, be less powerful than in the absence of measurement error. Note also that under the nondifferential measurement error assumption,

$$\mathrm{E}(Y|W, Z) = \beta^*_0 + \beta^*_1\mathrm{E}(X|W, Z) + \beta^*_2 Z.$$

It follows that, when $E(X|W, Z) = \alpha^*_0 + \alpha^*_1 W + \alpha^*_2 Z$,

$$\mathrm{E}(Y|W, Z) = \beta^*_0 + \beta^*_1\alpha^*_0 + \beta^*_1\alpha^*_1 W + (\beta^*_2 + \beta^*_1\alpha^*_2)Z. \tag{1.12}$$

This implies that the naive model test that none of the predictors are useful for explaining variation in $Y$ is valid in the sense of having the desired Type I error rate. Specifically, examination of (1.3) and (1.12) shows that $\theta^*_2 = 0$ is equivalent to $\beta^*_2 = 0$, only if $\beta^*_1\alpha^*_2 = 0$. It follows that the naive test of $H_0 : \beta^*_2 = 0$ is valid only if $X$ is unrelated to $Y$ conditional on $Z$ ($\beta^*_1 = 0$) or if $Z$ is unrelated to $X$ ($\alpha^*_2 = 0$). The naive tests that are valid, that is, those that maintain the Type I error rate, will still suffer reduced power relative to the test based on the true data.

## 4.2 Bias with the Berkson error model

The Berkson models the condition $\mathrm{E}(X|W) = W$. It follows from the fact that $W \perp\!\!\!\perp U$, $\mu_w = \mu_x$, $\sigma^2_x = \sigma^2_w + \sigma^2_u$ and

$$Cov(W, X) = Cov(W, W + U) = \sigma^2_w.$$

This implies, using (1.7), that $\lambda = 1$. That is, when the error in $X$ follows the unbiased Berkson error model, the naive estimator of slope is an unbiased estimator of $\beta^*$ in linear regression model. However, there is no bias in the naive regression parameter estimators, but there is an increase in the residual variance because under model (1.2) and the Berkson error model,

$$\begin{aligned} Var(Y|W) &= Var(\beta_0^* + \beta_1^* X + \epsilon|W) \\ &= \beta_1^{*2} Var(W + U|W) + \sigma_\epsilon^2 \\ &= \beta_1^{*2}\sigma_u^2 + \sigma_\epsilon^2. \end{aligned}$$

## 4.3 Nonlinear regression models

Regression coefficients in generalized linear models, including models of particular interest in epidemiology such as logistic regression (or probit regression) and poisson regression, are affected by measurement error in much the same manner as are linear model regression coefficients. In particular, relative risks and odds ratios are affected by measurement error much the same as linear model regression coefficients (Rosner et al. 1989, 1990; Stefanski 1985; Carroll et al., 2006). Suppose that the outcome variable $Y$ is dichotomous. We now consider model (1.1) with logistic link without covariates $Z$,

$$\text{logit}P(Y = 1|X) = \beta_0^* + \beta_1^* X. \tag{1.13}$$

Assume that the error in $X$ is nondifferential and follows the classical measurement error model. Then the observed data model implied by these restrictions, satisfies

$$P(Y = 1|W) = \int_x P(Y = 1|W, X) f_{X|W}(x|w)dx = \int_x P(Y = 1|X) f_{X|W}(x|w)dx.$$

This integral is not easy to handle, and to the best of our knowledge there is no closed form solution for the bias expressions. Now consider the above model with the probit link,

$$\Phi^{-1}\{P(Y = 1|X)\} = \beta_0^* + \beta_1^* X.$$

Under the assumption of normality, we can evaluate the latter integral by the probit link. Because, as stated, $X \sim N(\mu_x, \sigma_x^2)$ and $U \sim N(0, \sigma_u^2)$ then

$$\text{E}(X|W) = \mu_x + \rho_{xw}\frac{\sigma_x}{\sigma_w}(W - \mu_w)$$

$$Var(X|W) = (1 - \rho_{xw}^2)\sigma_x^2 = \left(\frac{1}{\sigma_x^2} + \frac{1}{\sigma_u^2}\right)^{-1}$$

because $\rho_{xw} = \frac{\sigma_x}{\sigma_w}$ in the classical error model. Then the latter integral can be written as

$$\begin{aligned} P(Y_i = 1|W_i) &= \int_{-\infty}^{\infty} \Phi(\beta_0^* + \beta_1^* x_i) f(x_i; \mu_i^*, \sigma^{*2}) dx_i \\ &= \Phi\left\{ \frac{\beta_0^* + \beta_1^* \mu_i^*}{\sqrt{1 + \beta_1^{*2} \sigma^{*2}}} \right\} \end{aligned}$$

where $\sigma^{*2} = (\frac{1}{\sigma_x^2} + \frac{1}{\sigma_u^2})^{-1}$ and $\mu_i^* = \mathrm{E}(X_i|W_i)$ for $i = 1, ..., n$ (Reeves et al., 1998; Heid et al., 2002). A direct comparison of the latter with the naive model

$$P(Y_i = 1|W_i) = \Phi(\theta_0^* + \theta_1^* W_i)$$

yields

$$\theta_1^* = \frac{\lambda \beta_1^*}{\sqrt{1 + \sigma_u^2 \lambda \beta_1^{*2}}}$$

and

$$\theta_0^* = \frac{\beta_0^* + (1 - \lambda)\mu_w \beta_1^*}{\sqrt{1 + \sigma_u^2 \lambda \beta_1^{*2}}}.$$

The close relationship between the logit and probit form, namely $G(t) = (1+\exp(-t))^{-1} \approx \Phi(t/h)$ with $h = 1.70$, allows us to obtain an approximate asymptotic bias formula for the logistic regression coefficients in model (1.13).

# 5 Discrete explanatory variables

As explained in Section (2.3), the situation in which a discrete variable is measured with error, is referred to as misclassification. The degree of misclassification error in $X$ can be expressed in terms of misclassification probabilities. In the case of a dichotomous explanatory variable $X$, the probability $\pi_{1|1,z} = P(W = 1|X = 1, Z)$, for instance expresses how likely it is for someone who is truly exposed with covariate level $Z$ to be classified as exposed. Likewise, $\pi_{0|0,z} = P(W = 0|X = 0, Z)$ expresses how likely it is for someone who is truly unexposed with covariate level $Z$, to be classified as unexposed. In view of this, for a dichotomous explanatory variable $X$ taking the values 0 and 1, the probabilities $\pi_{1|1,z}$ and $\pi_{0|0,z}$ are called sensitivity and specificity respectively. The extent to which $\pi_{1|1,z}$ and $\pi_{0|0,z}$ are less than 1 reflects the severity of the degree of misclassification, with 1 indicating no misclassification error. An alternative is to use reclassification probabilities, that is $P(X = x|W =$

$w, Z$). Spiegelman et al. (2000) use these to develop inference for logistic regression with covariate misclassification, and Christopher and Kupper (1995) refer to these as predictive classification probabilities. We adopt misclassification probabilities here and reclassification probabilities in chapter 5. In epidemiology, most discussions on the effects of misclassification of explanatory variables have focused on the impact on the relative risk or the odds ratio in studies of a dichotomous explanatory variable (e.g., Morrissey and Spiegelman, 1999; Gustafson, 2003; Höfler, 2005; Jurek et al., 2005). For instance, let $Y = 1$ indicate the presence of a particular disease and $X = 1$ indicate exposure to a putative risk. Typically, the inferential interest focuses on the odds ratio,

$$\psi = \frac{P(X = 1|Y = 1)/P(X = 0|Y = 1)}{P(X = 1|Y = 0)/P(X = 0|Y = 0)},$$

which describes the association between exposure and disease. Here, $P(X = 1|Y = 1)$ and $P(X = 1|Y = 0)$ are the prevalences of exposure amongst diseased and disease-free subjects respectively. For instance, Jurek et al. (2005) show how often an observed relative risk is an overestimate of the true relative risk when the bias is towards the null. In addition to cause bias in effect estimates of the odds ratio, misclassification error may lead to an exaggerated precision in confidence intervals. The reason is that misclassification errors may add further noise to the data (Reade-Christopher and Kupper, 1991; Neuhaus, 1999). Consider now a linear regression model (1.2) for a continuous outcome $Y$ given a dichotomous explanatory variable $X$ which is subject to misclassification. For simplicity, we first start with no error-free covariate $Z$. Assuming nondifferential error

$$\begin{aligned}
\mathrm{E}(Y|W) &= \mathrm{E}\{\mathrm{E}(Y|X)|W\} \\
&= \beta_0^* + \beta_1^* P(X = 1|W) \\
&= \beta_0^* + \beta_1^* \{W P(X = 1|W = 1) + (1 - W)P(X = 1|W = 0)\} \\
&= \beta_0^* + \beta_1^* \frac{\pi_{0|1}\mu_x}{1 - \mu_w} + \beta_1^* \left\{ \frac{\pi_{1|1}\mu_x}{\mu_w} + \frac{\pi_{0|0}(1 - \mu_x)}{1 - \mu_w} - 1 \right\} W
\end{aligned}$$

where $\pi_{w|x} = P(W = w|X = x)$ for $w = 0, 1$ and $x = 0, 1$. A direct comparison of the latter with the naive model, $\mathrm{E}(Y|W) = \theta_0^* + \theta_1^* W$, shows that

$$\theta_1^* = \frac{\mu_x(1 - \mu_x)}{\mu_w(1 - \mu_w)}(\pi_{1|1} + \pi_{0|0} - 1)\beta_1^* \tag{1.14}$$

and

$$\theta_0^* = \beta_0^* + \frac{(1 - \pi_{1|1})\mu_x}{1 - \mu_w}\beta_1^*.$$

Let $\kappa = \frac{\mu_x(1-\mu_x)}{\mu_w(1-\mu_w)}(\pi_{1|1} + \pi_{0|0} - 1)$. Note that $\mu_w = P(W = 1)$ is the function of $(\pi_{1|1}, \pi_{0|0}, \mu_x)$ because

$$P(W = 1) = \sum_{x=0}^{1} P(W = 1|X = x)P(X = x) = 1 - \pi_{0|0} + (\pi_{1|1} + \pi_{0|0} - 1)\mu_x.$$

Substituting the latter into (1.14) yields

$$\kappa = \frac{\mu_x(1-\mu_x)}{\{1 - \pi_{0|0} + (\pi_{1|1} + \pi_{0|0} - 1)\mu_x\}\{\pi_{0|0} - (\pi_{1|1} + \pi_{0|0} - 1)\mu_x\}}(\pi_{1|1} + \pi_{0|0} - 1).$$

As in the continuous case, $\kappa$ is often referred to as attenuation factor. When there is no misclassification, that is; $\pi_{1|1} = \pi_{0|0} = 1$, then evidently $\theta_1^* = \beta_1^*$. To interpret $\kappa$, we take its derivative, for instance, with respect to $\pi_{1|1}$ for fixed $\pi_{0|0}$ and $\mu_x$:

$$\partial\kappa/\partial\pi_{1|1} = \mu_x(1-\mu_x)\frac{a(1-a) - (2a-1)(\pi_{0|0} - a)}{\{a(1-a)\}^2} \geq 0$$

where $a = \pi_{0|0} - (\pi_{1|1} + \pi_{0|0} - 1)\mu_x$. This implies that $\kappa$ is increasing in $\pi_{1|1}$ for fixed $\pi_{0|0}$ and $\mu_w$. Likewise $\kappa$ is increasing in $\pi_{0|0}$ for fixed $\pi_{1|1}$ and $\mu_x$. Thus the effect of misclassification is an attenuation bias. As expected, the bias also worsens with the severity of the misclassification. It is usually expected that nondifferential misclassification of a dichotomous explanatory variable will yield bias towards the null, although this rule can break down when the variable is polytomous (Gustafson, 2003; Jurek et al., 2005; Chu et al., 2006). Further, When $\mu_x = 0.5$, the bias is symmetric in $\pi_{1|1}$ and $\pi_{0|0}$, as can be seen by direct inspection of (1.14). For instance, when $\mu_x = 0.5$ and $\pi_{1|1} = \pi_{0|0} = 0.9$ (which can be interpreted as 10% misclassification), then $\theta_1^* = 0.80\beta_1^*$, because $\mu_w = 0.45$ and $\kappa = 0.80$. This can be interpreted as 20% attenuation. Note that, as for continuous explanatory variables, tests whether the error-prone explanatory variable is associated with the outcome, remain valid in the presence of measurement error because the bias expression (1.14) is zero under the null hypothesis $H_0$: $\beta_1^* = 0$. They may however be less powerful than in the absence of measurement error. Further, when $\kappa$ is known then $\mathrm{E}(\hat{\beta}_1) = \kappa^{-1}\mathrm{E}(\hat{\theta}_1) = \beta_1^*$ and $Var(\hat{\beta}_1) = \kappa^{-2}Var(\hat{\theta}_1)$. More generally when there is an error-free covariate $Z$ in model (1.2) with a dichotomous explanatory variable $X$, misclassification error may be expressed in terms of

$$P(W = 1|X, Z = z) = 1 - \pi_{0|0,z} + (\pi_{1|1,z} + \pi_{0|0,z} - 1)X,$$

where $\pi_{w|x,z}$ is related with the error-free covariate $Z$ for $w = 0, 1$ and $x = 0, 1$. In chapter 5, we investigate the asymptotic bias of the ordinary least squares estimate

of the regression coefficient in terms of reclassification probabilities when they are related to the error-free covariate $Z$. We now investigate the bias of the ordinary least squares estimate of the regression coefficients $\beta^*$ when

$$\pi_{w|x,z} = \pi_{w|x}$$

for $w = 0, 1$ and $x = 0, 1$. We show by solving the following expected estimating equation under the nondifferential measurement error,

$$\begin{aligned} 0 = \mathrm{E}\{U(\theta^*)\} &= \mathrm{E}\left\{ \begin{pmatrix} 1 \\ W \\ Z \end{pmatrix} (Y - \theta_0^* - \theta_1^* W - \theta_2^* Z) \right\} \\ &= \mathrm{E}\left\{ \begin{pmatrix} 1 \\ W \\ Z \end{pmatrix} (\mathrm{E}(Y|X,Z) - \theta_0^* - \theta_1^* W - \theta_2^* Z) \right\} \end{aligned}$$

that $\theta_0^* = \beta_0^* + \mu_x \beta_1^* - \mu_w \theta_1^* + \mu_z(\theta_2^* - \beta_2^*)$.
This can obviously be obtained by solving the first row of the above equation. In the Appendix, we further show that

$$\theta_1^* = (\pi_{1|1} + \pi_{0|0} - 1) \left\{ \frac{\mu_x(1-\mu_x)(1-\rho^2)}{\mu_w(1-\mu_w) - \mu_x(1-\mu_x)(\pi_{1|1} + \pi_{0|0} - 1)^2\rho^2} \right\} \beta_1^* \qquad (1.15)$$

and

$$\theta_2^* = \beta_2^* + \beta_1^* \rho \frac{\sqrt{\mu_x(1-\mu_x)}}{\sigma_z} \left\{ 1 - (\pi_{1|1} + \pi_{0|0} - 1) \frac{\theta_1^*}{\beta_1^*} \right\}$$

where $\rho = \rho_{xz}$ and $\sigma_z = \sqrt{Var(Z)}$. The coefficient

$$\kappa_1 = (\pi_{1|1} + \pi_{0|0} - 1) \left\{ \frac{\mu_x(1-\mu_x)(1-\rho^2)}{\mu_w(1-\mu_w) - \mu_x(1-\mu_x)(\pi_{1|1} + \pi_{0|0} - 1)^2\rho^2} \right\}$$

is again referred to as the attenuation factor. By taking partial derivatives of $\kappa_1$, it is straightforward to show that it is increasing in $\pi_{1|1}$ for fixed $(\pi_{0|0}, \mu_x, \rho)$ and increasing in $\pi_{0|0}$ for the fixed $(\pi_{1|1}, \mu_x, \rho)$. For the fixed value of $(\pi_{1|1}, \pi_{0|0}, \mu_x)$, $\kappa_1$ is decreasing in $|\rho|$ (Gustafson, 2003). The latter implies that the attenuation worsens as the correlation between $X$ and $Z$ increases.

# 6 Methods for measurement error correction

As stated before, an important problem in most measurement error analysis is the inability to correct for bias due to measurement error given only the information contained in the sample of observed $(Y, W, Z)$ variables. That is, one would not

be able to correct for bias due to measurement error only based on the information contained in $(Y, W, Z)$. Over the past decades, a number of statistical techniques have been proposed for correcting the impact of measurement error and misclassification error in an explanatory variable $X$. The choice of methods depends on the information distribution of the variables, the magnitude of the error variance, and the type of error model. They indeed differ according to the assumptions about the distribution of the unobserved explanatory variable $X$, the availability of additional data about the unobserved explanatory variable $X$ and the theoretical background of the approach, which may be parametric or nonparametric. Fuller (1987) developed extensive methods for measurement error correction in linear regression models. A review of measurement error correction techniques in case-control studies is given in Thürigen el al. (2000). Gustafson (2003) has presented correction methods with a focus on Bayesian adjustments. Carroll et al. (2006) have provided a comprehensive account of current statistical methodologies for measurement error correction in nonlinear regression models.
Because under nondifferential measurement error, the observed data likelihood equals

$$\begin{aligned} f(Y, W, Z) &= \int f(Y|X, Z) f(W|X, Z) f(X, Z) dX \\ or &= \int f(Y|X, Z) f(X|W, Z) f(W, Z) dX \end{aligned}$$

measurement error correction requires information about either the distribution of $W$ given $X$ (when the error in $X$ follows the classical measurement error or misclassification probabilities) or of $X$ given $W$ (when the error in $X$ follows the Berkson error model or reclassification probabilities) possibly conditional on $Z$. Because $X$ is unobserved, one usually needs to observe additional data sources, beyond the main study sample. Additional data can be available in different forms. For instance, a subsample of observations from $X$ can be recorded for a small group of subjects of the main study sample. It yields an internal validation data set, from which so-called gold standard measures of $X$ are available. It follows naturally that missing data methods can be applied for measurement error correction in this setting (Robins, Rotnitzky and Zhao, 1994; Carroll et al., 2006). A common alternative is to collect replication data; that is, replicates $\{W_{ij}; j = 1, ..., k_i\}$ of the observations for $W_i$. The simple case is where $k_i = 2$. In this case, one may choose $W_i = (W_{i1} + W_{i2})/2$ as an observed explanatory variable, because the average of the $W_{ij}$ is a better estimate of $X_i$ than $W_i$ alone. For example, the average ancestry proportion computed on a set of full siblings would be a more accurate measure of their ancestry proportion than the value observed on a single individual (Divers et al., 2007). "The individual admixture proportion estimates obtained by using ancestry informative markers are measure-

ments of the underlying individual ancestry proportion". Such measurements may also help to estimate the measurement error variance $\sigma_u^2$. Suppose that the classical error model holds and that $k_i$ replicate measurements of $X_i$ are available. Then the usual component of variance analysis (Carroll and Stefanski, 1990; Wang et al., 1995; Carroll et al., 2006) yields the following estimate of the measurement error variance on $W_{ij}$

$$\hat{\sigma}_u^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k_i}(W_{ij} - \bar{W}_{i0})(W_{ij} - \bar{W}_{i0})^t}{\sum_{i=1}^{n}(k_i - 1)} \tag{1.16}$$

where $\bar{W}_{i0} = \sum_{j=1}^{k_i} W_{ij}/k_i$. We will illustrate how to estimate $\sigma_u^2$ on Framingham Heart study in the next sections. In the absence of validation data or replicates, sometimes an (unbiased) instrument data $T_i$ may be available for a subset of the study participants. Instrument data are another measurements of $X$ in addition to $W$. Instrument data or instrumental variable is associated with $X$, independent of measurement error $U$, and has no information about the outcome $Y$ other than what is available in $X$ (see Section 6.3). Thus, there are generally three kinds of additional data to that guarantee the parameter identifiability in the present of measurement error

- Validation data in which $X$ is observed directly;
- Replication data, in which replicates of $W$ are available;
- Instrument data, in which another variable $T$ is observable in addition to $W$.

In the next subsections, we illustrate methods for correcting for bias due to measurement error or misclassification error in regression models.

## 6.1 Regression calibration

Regression calibration is a straightforward approach to correct for bias due to measurement error and has been successfully applied to a broad range of regression models, in particular linear and logistic regression (Rosner et al., 1996; Thürigen et al., 2000; Buzas et al., 2005; Spiegelman et al., 2005; Carroll et al., 2006). The idea underlying this method is the estimation of the regression of $X$ on $W$, possibly adjusted for error-free covariates $Z$. That is, first one obtains an estimate of the expected explanatory variable $X$ in function of $W$ and $Z$ by fitting an appropriate

regression (often linear) model, for instance, to an validation data. Then, one fits a regression of $Y$ on $\hat{X}$ and $Z$, rather than on $X$ and $Z$, to obtain parameter estimates, where $\hat{X} \equiv \mathrm{E}(X|W, Z; \hat{\alpha})$, and $\hat{\alpha}$ is the estimated regression parameter of $X$ on $W$ and $Z$. This methodology can be seen in (1.3) and (1.4). Note that for dichotomous outcome, (1.4) can be approximately written

$$\mathrm{logit}P(Y = 1|W, Z) \approx \beta_0^* + \beta_1^* \mathrm{E}(X|W, Z) + \beta_2^* Z$$

where $\mathrm{logit}(p) = \log\{p/(1-p)\}$. This suggests the following two-stage approach:

1. Using validation data, replicate data or instrument data to fit a regression of $X$ on $W$ and $Z$ to obtain $\hat{X}$;
2. Replace the unobserved $X$ by $\hat{X}$ in the regression of $Y$ on $X$ and $Z$.

Standard errors of the parameters estimates obtained in step 2 must account for the fact that $\hat{X}$ is estimated in step 1 and can be obtained using either the bootstrap or a sandwich method (e.g., Carroll et al., 2006).
The best linear approximation to $X$ given $(W, Z)$ is

$$\mathrm{E}(X|W, Z) \approx \mu_x + \frac{\sigma_{wz}\sigma_{xz} - \sigma_{wx}\sigma_z^2}{\sigma_{wz}^2 - \sigma_w^2\sigma_z^2}(W - \mu_w) + \frac{\sigma_w^2\sigma_{xz} - \sigma_{wx}}{\sigma_w^2\sigma_z^2 - \sigma_{wz}^2}(Z - \mu_z).$$

The coefficients are calculated from the linear regression $X$ on $(W, Z)$. This approximation is exact when $X$, $W$ and $Z$ are jointly normally distributed. When the classical error model holds, the latter approximation is equal to

$$\mathrm{E}(X|W, Z) \approx \mu_w + \begin{pmatrix} \sigma_x^2 \\ \sigma_{xz} \end{pmatrix}' \begin{bmatrix} \sigma_x^2 + \sigma_u^2 & \sigma_{xz} \\ \sigma_{xz} & \sigma_z^2 \end{bmatrix}^{-1} \begin{pmatrix} W - \mu_w \\ Z - \mu_z \end{pmatrix} \tag{1.17}$$

because $\mu_x = \mu_w$, $\sigma_{wx} = \sigma_x^2$, and $\sigma_{wz} = \sigma_{xz}$. Note from (1.17) that, when there is no validation data, the relationship between $X$ and $(W, Z)$ may be inferred from replicate data or an (unbiased) instrument $T$. For replicates data, for instance with $k_i = 2$ and $W_i = (W_{i1} + W_{i2})/2$, this follows with

$$\hat{\sigma}_u^2 = \frac{\sum_{i=1}^n (W_{i1} - W_i)^2 + \sum_{i=1}^n (W_{i2} - W_i)^2}{n}$$

and

$$\hat{\sigma}_w^2 = (n-1)^{-1} \sum_{i=1}^n (W_i - \hat{\mu}_w)^2.$$

For instrument data, this is because, by definition of unbiased instrument $\mathrm{E}(T|W, Z) = \mathrm{E}(X|W, Z)$ (Rosner et al., 1990; Carroll et al., 2006).

## Application: analysis of the Framingham study

The Framingham study on coronary heart disease and blood pressure is a large cohort study consisting of $n = 1615$ men aged 31-65 years. The outcome, $Y$, indicates 1 for the occurrence of coronary heart disease (CHD) within an eight-year period following exam 3 and 0 otherwise. There are a series of exams taken two years apart which exam 3 uses as the baseline. Predictors employed in this example are the patient's age at exam 2, smoking status at exam 1, serum cholesterol at exam 2 and 3, and systolic blood pressure (SBP) at exam 2 and 3, rather as the average of two measurements taken by different examiners during the same visit. In this data set, we consider age, smoking status, and serum cholesterol at exam 3 (cho3) as the error-free covariates $Z$. The interest lies in fitting a logistic regression model of CHD on systolic blood pressure (SBP) which are measured with error conditional on the error-free covariate $Z$. The observed explanatory variable $W$ is a modified version of a transformation of SBP, that is, $W = \log(SBP - 50)$ as suggested in Spiegelman et al. (1984) and Carroll et al. (2006) to make the normality assumption for the measurement error more plausible. The unobserved explanatory variable $X$ is defined to be the long-term average of $W$. Specifically, $SBP_{mj}$ indicates the $j$th measurement of SBP from the $m$th exam, $j = 1, 2$, $m = 2, 3$. The transformed $SBP$ are

$$W_{i1} = \log\{(SBP_{i31} + SBP_{i32})/2 - 50\}$$

and

$$W_{i2} = \log\{(SBP_{i21} + SBP_{i22})/2 - 50\}$$

for each subject $i = 1, ..., 1615$. The overall surrogate is $\bar{W}_{i0} = (W_{i1} + W_{i2})/2$, the sample mean for each subject $i$. The error model is $W = X + U$, where $U$ is assumed to be independent of $X$ with mean zero and variance $\sigma_u^2$.

**Example 6**. We now perform a measurement error analysis for the Framingham data by the regression calibration approach. Validation data are unavailable in this study; instead, two replicates of $X$ are available. We will therefore rely on replicate data. Let $W_i = \bar{W}_{i0} = (W_{i1} + W_{i2})/2$. Then $\hat{\mu}_w = \sum_{i=1}^{n} \bar{W}_{i0}/n = 4.364$ and $\hat{\sigma}_w^2 = 0.045$. It follows from (1.16) for $k_i = 2$ that

$$\hat{\sigma}_u^2 = \frac{\sum_{i=1}^{n}(W_{i1} - \bar{W}_{i0})^2 + \sum_{i=1}^{n}(W_{i2} - \bar{W}_{i0})^2}{n} = 0.013.$$

Note that, the measurement error variance on $W_i$ is only half of this, namely 0.006.

It follows from (1.17) that:

$$\hat{\mathrm{E}}(X_i|W_i, Z_i) \approx 4.364 + \begin{pmatrix} 0.039 \\ \hat{\sum}_{xz} \end{pmatrix}^t \begin{bmatrix} 0.045 & \hat{\sum}_{xz} \\ \hat{\sum}_{xz}^t & \hat{\sum}_{zz} \end{bmatrix}^{-1} \begin{pmatrix} W_i - 4.364 \\ Z_i - \hat{\mu}_z \end{pmatrix}$$

where $\hat{\sigma}_x^2 = \hat{\sigma}_w^2 - \hat{\sigma}_u^2 = 0.045 - 0.006 = 0.039$, $Z = (age, smoke, cho3)$ and $\hat{\mu}_z = (45.86, 0.77, 228.40)$. For $k_i = 2$;

$$\hat{\sum}_{xz} = (n-1)^{-1} \sum_{i=1}^{n} (W_i - \hat{\mu}_w)(Z_i - \hat{\mu}_z)^t$$

$$\hat{\sum}_{zz} = (n-1)^{-1} \sum_{i=1}^{n} (Z_i - \hat{\mu}_z)(Z_i - \hat{\mu}_z)^t$$

where $\hat{\sum}_{xz}$ is the sample covariance of $(X, Z)$ and $\hat{\sum}_{zz}$ is the sample covariance of $Z$. Remember that, when measurement error follows the classical error model, $Cov(W, Z) = Cov(X, Z)$. We then fit a logistic regression of $Y$ on $\hat{X} = \mathrm{E}(X|W, Z; \hat{\alpha})$ and $Z$. The naive estimates are

$$\mathrm{logit}P(Y = 1|W, Z) = -14.949 + 1.706\ W + 0.055\ age + 0.593\ smoke + 0.008 cho3$$

and the regression calibration estimates are

$$\mathrm{logit}P(Y = 1|\hat{X}, Z) = -16.180 + 2.013\ \hat{X} + 0.053\ age + 0.601\ smoke + 0.008 cho3$$

When we perform the regression calibration approach with $W_i = W_{i1}$, so that $ki = 1$. Then $\hat{\sigma}_u^2 = 0.013$, because this is variance measurement error either on $W_{i1}$ or on $W_{i1}$. In this case, $\hat{\mu}_w = 4.355$ and $\hat{\sigma}_w^2 = 0.052$. Now the naive estimates are

$$\mathrm{logit}P(Y = 1|W, Z) = -14.182 + 1.524\ W + 0.057\ age + 0.573\ smoke + 0.008 cho3$$

and the regression calibration estimates are

$$\mathrm{logit}P(Y = 1|\hat{X}, Z) = -16.474 + 2.099\ \hat{X} + 0.053\ age + 0.582\ smoke + 0.007 cho3$$

Table 1.1 summarizes the naive and corrected slope estimate by regression calibration along with the standard errors when the error-free covariates are age, smoking, cholesterol level. Standard errors for the regression calibration estimators are obtained by bootstrap. As Table 1.1 shows, the corrected estimator has bigger standard error than the naive estimator. This is the price to pay for reduced bias. The results also

show when $W$ is considered to be the average systolic blood pressure at Exams 2, 3, the regression calibration estimator is is more efficient ($SE = 0.471$) than when it is considered only at Exam 3 ($SE = 0.523$). This is because by selecting the average of SBP3 and SBP2, the measurement error variance is twice smaller than by selecting only SBP3.

Table 1.1: *Regression calibration method to correct measurement error on Framingham study. The Naive and the regression calibration estimator and their standard error are presented. Standard errors for regression calibration estimator are obtained based on 10 000 bootstrap resamples.*

| Estimate | W | $\hat{\beta}_1$ | SE | $\hat{\sigma}_u^2$ |
|---|---|---|---|---|
| Naive | SBP3 | 1.524 | 0.389 | - |
| | (SBP2+SBP3)/2 | 1.706 | 0.417 | - |
| R.C. | SBP3 | 2.070 | 0.523 | 0.013 |
| | (SBP2+SBP3)/2 | 2.011 | 0.471 | 0.006 |

## 6.2 Simulation-extrapolation

Simulation-extrapolation (SIMEX) is a useful tool for correcting measurement error in a very broad range of settings. This is the only method that provides a visual display of the impact of measurement error on regression parameter estimations (Cook and Stefanski, 1994; Stefanski and Cook, 1995 ; Carroll et al., 1996; Küchenhoff et al., 2006; Carroll et al., 2006). This method was originally proposed by Cook and Stefanski (1994) for parametric measurement error models in which the measurement error variance is known or at least well estimated. This method is helpful for complex models with a simple measurement error model (Küchenhoff et al., 2006). The basic idea underlying SIMEX is the fact that the impact of measurement error on regression coefficients can be determined through simulation. The effect of measurement error can next be eliminated by extrapolation. In fact, through SIMEX, new samples with larger error variances are generated by adding simulated errors to the original observed regressor variables. In this section, we briefly describe the SIMEX approach and illustrate it on the Framingham study.
The SIMEX procedure is generally performed in two steps:

1. Simulation: add additional measurement error in known increments to the data and compute estimates from the contaminated data. Next, establish a trend between the estimates and the variance of the added errors;

2. Extrapolation: Extrapolate this trend back to the case of no measurement error.

Consider for simplicity the linear regression model $\mathrm{E}(Y|X) = \beta_0^* + \beta_1^* X$ with a classical error model. Suppose that in addition to the $(Y, X, W)$ data used to estimate the naive coefficient $\theta_1^*$ in model $\mathrm{E}(Y|W) = \theta_0^* + \theta_1^* W$, there are $M-1$ additional data sets available, each with large error variance of $(1+\lambda_m)\sigma_u^2$, where $0 = \lambda_1 < \lambda_2 < ... < \lambda_M$ and $m = 1, ..., M$. That is, assuming that the variance of the measurement error $\sigma_u^2$ is known, one simulates

$$W_{ib}(\lambda) = W_i + \sqrt{\lambda} U_{ib} \quad i = 1, ..., n, \quad b = 1, ..., B$$

where $U_{ib}$ are independent standard normally distributed with mean 0 and variance $\sigma_u^2$. The least squares estimate of the slope from the $m$th data set, $\hat{\theta}_{1m}$, consistently estimates $\frac{\sigma_x^2}{\sigma_x^2 + (1+\lambda_m)\sigma_u^2}\beta_1^*$ (Cook and Stefanski, 1994; Carroll et al., 2006). The dependent variables $\{\hat{\theta}_{1m}; m = 1, ..., M\}$ thus relate nonlinearly to the independent variables $\{\lambda_m; m = 1, ..., M\}$ with mean function

$$\Gamma(\lambda) = \frac{\sigma_x^2}{\sigma_x^2 + (1+\lambda)\sigma_u^2}\beta_1^* \qquad \lambda \geq 0.$$

The parameter of interest, $\beta_1^*$ can thus be obtained from $\Gamma(\lambda)$ by extrapolation to $\lambda = -1$; the naive estimate occurs at $\lambda = 0$. Specifically, let $\hat{\theta}_s(\lambda_m)$ denote the vector of regression parameter estimators obtained by regression of $Y$ on $W_b(\lambda_m)$ for $m = 1, ..., M$ and $\hat{\theta}(\lambda_m) = B^{-1} \sum_{s=1}^{B} \hat{\theta}_s(\lambda_m)$. Here taking the average over the $B$ simulated data sets is needed to eliminate simulation error. Empirical evidence suggests that $B = 100$ is sufficient. For each value of $\lambda$, the parameters $\Theta(\lambda) = (\beta_0(\lambda), \beta_1(\lambda))$ corresponded to the $(\beta_0^*, \beta_1^*)$ and their corresponding standard errors are estimated $B$ times using a chosen estimation method (ordinary least squares, quadratic, nonlinear, etc.). In the second step, each component of the vector $\hat{\theta}(\lambda)$ is then modeled as a function of $\lambda$ and the SIMEX estimator is the extrapolation to $\lambda = -1$, which corresponds the ideal case of no measurement error. When the measurement error variance $\sigma_u^2$ is unknown, it must be estimated with extra data and then substituted $\sigma_u^2$ in the SIMEX approach. The unbiasedness of SIMEX estimator depends crucially on knowing $\sigma_u^2$ to being able to determine the functional form of $\hat{\theta}(\lambda)$, and on normality of the error. For variance estimation, three methods are available: the Delta method

(Carroll et al., 1996), Jackknife type estimators (Stefanski and Cook, 1995), and the Bootstrap.

Table 1.2: *Simulation-extrapolation approach to correct for measurement error in the Framingham study. The first row is when W is considered to be systolic blood pressure at Exam 3. The second row considers W to be the average systolic blood pressure at Exams 2, 3. The Naive and corrected estimators with their standard error are presented. The standard errors are obtained via Jackknife variance estimator.*

| Estimate | W | $\hat{\beta}_1$ | SE | $\hat{\sigma}_u^2$ |
|---|---|---|---|---|
| Naive | SBP3 | 1.524 | 0.389 | - |
| | (SBP2+SBP3)/2 | 1.706 | 0.417 | - |
| SIMEX | SBP3 | 1.865 | 0.473 | 0.013 |
| | (SBP2+SBP3)/2 | 1.931 | 0.462 | 0.006 |

We will perform the SIMEX approach on Framingham study in next example. In this example, we use the R-package *simex* which is written by Wolfgang Lederer (Lederer and Küchenhoff, 2006).

**Example 7**. We use the replicate SBP measurement from exam 2 and exam 3 for all study participants. The error model is the classical error model. Like in example 6, we perform this approach for two cases.

- With $W_i = \bar{W}_{i0} = (W_{i1} + W_{i2})/2$. The naive model is

  $$\text{logit}P(Y = 1|W, Z) = -14.949 + 1.706\, W + 0.055\, age + 0.593\, smoke + 0.007\, cho3$$

  The SIMEX procedure is repeated $B = 1000$ times for $\lambda = (0.5, 1, 1.5, 2)$. The resulting naive estimator and SIMEX corrected estimator are obtained. Table 2 shows the results for the naive and SIMEX corrected estimator with their standard error. Figure 2 contains plots of the logistic regression coefficients by SIMEX (solid circles). The points plotted at $\lambda = 0$ are the naive estimates.

- We now choose $W_i = W_{i1}$, then the naive model is

  $$\text{logit}P(Y = 1|W, Z) = -14.182 + 1.524\, W + 0.057\, age + 0.573\, smoke + 0.008\, cho3$$

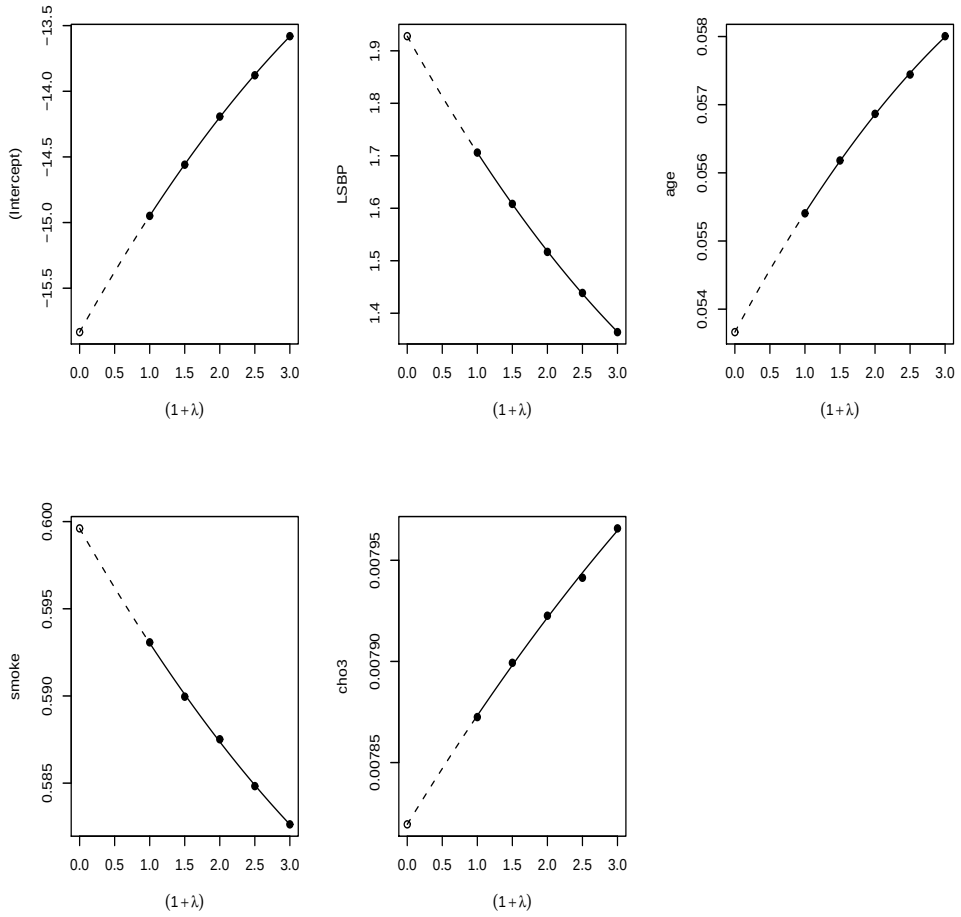



Figure 2: *Coefficient extrapolation for the Framingham logistic regression modeling. The simulated estimates are plotted (solid circles) and extrapolated to* $\lambda = 0$ *(dashed line) resulting in the SIMEX estimate. Open circles indicates SIMEX estimates obtained with the quadratic extrapolant.*

As we calculated in example 6, $\hat{\sigma}_u^2 = 0.013$, for $k_i = 1$. The SIMEX procedure is repeated $B = 1000$ times for $\lambda = (0.5, 1, 1.5, 2)$. Table 1.2 summarizes the results for the naive and SIMEX corrected estimator with their standard error. The standard error is obtained via Jackknife variance estimator through the *simex* package.

The standard error for the corrected estimators is bigger than for the naive estimator. This is the price to pay for having smaller bias.

## 6.3 Instrumental variable methods

In many situations in measurement error analysis, particularly in epidemiological contexts, neither validation measurements nor replication data can be obtained. One interesting strategy for measurement error analysis is then to select an instrumental variable (IV). This is a variable that is

- associated with $X$;
- independent of $W - X$ where $W = X + U$, that is, $T \perp\!\!\!\perp U$;
- conditionally independent of outcome $Y$ given $X$ and $Z$, that is, $T \perp\!\!\!\perp Y|X, Z$.

Instrumental variables (IV's) in the context of measurement error are often secondary measurements of the true explanatory variable $X$ other than $W$ which need not be unbiased for $X$. They are obtained by an independent methods (Carroll et al., 2006). For instance in agronomic experiment, $W$ could be the observed nitrogen in the leaves of the plant and $Y$ is the dry weight of the plant. We would then expect the true nitrogen in the leaves, $X$, to be correlated with nitrogen fertilizer, $T$, applied to the experimental plot (Fuller, 1987). Further, as agronomists believe the nitrogen fertilizer does not provide any more information about the dry weight of the plant than the nitrogen in the leaves and is independent of the measurement error on the nitrogen in the leaves. Then nitrogen fertilizer can be considered as an instrumental variable for the nitrogen in the leaves. Note that a replicate measurement of $X$ can be considered as an IV, because replicate measurements of $X$ are associated with $X$, and are independent of measurement error on $X$, and have no more information about $Y$ than $X$. Note also that an IV is not necessarily a replicate because $T$ is independently measured for $X$ (Buzas et al., 2005; Carroll et al., 2006). In the Framingham study, we can use systolic blood pressure at Exam 2, $W_2$, as an instrumental variable for correcting measurement error.
Roughly, IV is a variable which is highly correlated with the error-prone explanatory variable $X$, but is not associated with the outcome $Y$ other than through $X$ or $Z$,

and is independent of measurement error $U$. Assumption (1) is important because IV-estimators are generally effective only when $X$ and $T$ are strongly correlated. Assumption (2) says that $T$ does not dependent on the measurement error in $X$. Assumption (3) specifies that $T$ does not provide any more information about $Y$ than $X$. $T$ is therefore often called a surrogate. If an instrument or an instrumental variable (IV) is available, consistent estimates may still be obtained in the presence of measurement error (Fuller, 1987; Amemiya 1990; Stefanski and Carroll, 1991 ; Stefanski and Buzas, 1995; Buzas et al., 2005; Carroll et al., 2006; Lewbel, 2007). In fact, estimation method for IVs are widely recognized as an important method of analysis of linear measurement error models and the best-known method is two-stage least squares estimation. We first review linear model instrumental variable estimation in its simplest form. Suppose that $Y$ is linearly related to the error-prone explanatory variable $X$, $\mathrm{E}(Y|X) = \beta_0^* + \beta_1^* X$, and that the classical error model holds. The interest is to estimate $\beta_1^*$. Suppose that the relation between $X$ and $T$ is also linearly expressed as

$$\mathrm{E}(X|T) = \alpha_0^* + \alpha_1^* T.$$

Then with the classical error model

$$\mathrm{E}(W|T) = \mathrm{E}(X + U|T) = \alpha_0^* + \alpha_1^* T + \mathrm{E}(U|T).$$

It follows from assumption (2) that $\mathrm{E}(X|T) = \mathrm{E}(W|T)$. It also follows from the assumption (3) that

$$\begin{aligned} \mathrm{E}(Y|T) &= \mathrm{E}\{\mathrm{E}(Y|T, X)|T\} \\ &= \beta_0^* + \beta_1^* \mathrm{E}(X|T) \\ &= \beta_0^* + \beta_1^* \alpha_0^* + \beta_1^* \alpha_1^* T. \end{aligned}$$

That is the slope coefficient of regression $Y$ on $T$ is equal to the product of the slope coefficient of a regression $Y$ on $X$ and the slope coefficient of a regression $W$ on $T$. Note that under the assumption (1), $\alpha_1^*$ is nonzero. The IV-estimator can then be obtained by a two-stage least squares algorithm as follows:

1. Fit the linear regression of $W$ on $T$, $\mathrm{E}(W|T; \alpha^*)$, to find an estimate of $\hat{\alpha}$ of $\alpha^*$;
2. Fit the linear regression of $Y$ on the predicted values $\mathrm{E}(W|T; \hat{\alpha})$.

Equivalently, an IV-estimator for $\beta_1^*$ can be obtained as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(T_i - \bar{T})}{\sum_{i=1}^{n} (W_i - \bar{W})(T_i - \bar{T})}.$$

Fuller (1987) shows that this is a consistent estimator for $\beta_1^*$. A consistent estimator for $Var(\hat{\beta}_1)$ can be obtained by the sample moments

$$\hat{Var}(\hat{\beta}_1) = (n-1)^{-1} M_{wt}^{-2} M_{tt} S$$

where $M_{tt} = n^{-1} \sum_{i=1}^{n} (T_i - \bar{T})(T_i - \bar{T})^t$, $M_{wt} = n^{-1} \sum_{i=1}^{n} (W_i - \bar{W})(T_i - \bar{T})^t$, and $S = (n-2)^{-1} \sum_{i=1}^{n} \{Y_i - \bar{Y}) - \hat{\beta}_1 (W_i - \bar{W})\}^2$.

The most common solution to the measurement error analysis for the linear regression model is the use of instrumental variable estimation. This methodology cannot be easily applied in the nonlinear regression framework (Stefanski and Buzas, 1995; Carroll et al., 2006). However, the IV approach is the most widely used technique for dealing with error in explanatory variables in linear multiple regression problems. To date the methods proposed for the nonlinear model depend on very strong restrictions on the distribution of the measurement errors of the variables which correspond to the unknown regression coefficients (Hausman et al., 1995). Hausman et al. (1995) discuss consistent estimators for nonlinear (polynomial) regression specifications in which estimators depend on the existence of instrumental variables or a single repeated observation. Stefanski and Buzas (1995) describe two approaches to instrumental variable estimation in binary regression measurement error models. Their methods entail constructing approximated mean models for the binary outcome as a function of the measured predictor, the instrument and any covariates in the model. Instrumental variable estimation for generalized linear measurement error models, that includes linear and logistic regression as special cases, are considered by Buzas and Stefanski (1996). Carroll et al. (2006) describe the IV methods in a class of nonlinear measurement error in a way that is closely related to the regression calibration method. We now review the case where the outcome $Y$ is dichotomous. Suppose that

$$\text{logit} P(Y = 1|X) = \beta_0^* + \beta_1^* X.$$

It also follows from the assumptions of IV that

$$\begin{aligned} P(Y = 1|T) &= \text{E}\{P(Y = 1|T, X)|T\} \\ &= \text{E}\{\text{expit}(\beta_0^* + \beta_1^* X)|T\} \\ &\approx \text{expit}\{\beta_0^* + \beta_1^* \text{E}(X|T)\}. \end{aligned}$$

The last step is implied by $\mathrm{E}(X|T) = \mathrm{E}(W|T)$ and the Taylor series expansion

$$\begin{aligned}\mathrm{E}\left\{\mathrm{expit}(\beta_0^* + \beta_1^* X)|T\right\} &\approx \mathrm{E}\left[\mathrm{expit}\{\beta_0^* + \beta_1^* \mathrm{E}(X|T)\}\right. \\ &\quad \left. + \mathrm{expit}'\{\beta_0^* + \beta_1^* \mathrm{E}(X|T)\}\{X - \mathrm{E}(X|T)\} + ...|T\right] \\ &= \mathrm{expit}\{\beta_0^* + \beta_1^* \mathrm{E}(W|T)\}\end{aligned}$$

where $\mathrm{expit}(a) = \exp(a)/\{1 + \exp(a)\}$. This estimator can be obtained by a two-stage algorithm as follows

1. Fit the linear regression of $W$ on $T$, $\mathrm{E}(W|T;\alpha^*)$, to find an estimate of $\hat{\alpha}$ of $\alpha^*$;
2. Fit the logistic regression of $Y$ on the predicted values $\mathrm{E}(W|T;\hat{\alpha})$.

The resulting IV-estimator is consistent and an estimate of variance may be obtained by bootstrap or sandwich method. We now perform the IV methods for the Framingham study in the next example.

Table 1.3: *Results of the instrumental variable method to correct for measurement error on Framingham study. The instrumental variable is systolic blood pressure at exam 2. The naive and the instrumental variable estimator with their standard errors are presented. The IV-estimator's standard error is obtained based on 10 000 bootstrap resamples.*

| Estimate | IV | $\hat{\beta}_1$ | SE |
|---|---|---|---|
| Naive estimator | - | 1.524 | 0.389 |
| IV-estimator | SBP3 | 2.002 | 0.520 |

**Example 8**. Consider again the Framingham heart study, wherein two systolic blood pressure measurements from each of two exams were measured. We consider $W_i = W_{i1}$ and $T_i = W_{i2}$. First we fit a linear regression of $T$ on $W$ and $Z$:

$$\mathrm{E}(W|T, Z) = 0.935 + 0.754\ T + 0.001\ \mathit{age} + 0.012\ \mathit{smoke} + 0.0002\ \mathit{cho3}.$$

Then we fit a logistic regression of $Y$ on $\hat{T} = E(W|T, Z;\hat{\alpha})$ where $\hat{\alpha}$ was obtained in the first step:

$$\mathrm{logit}P(Y = 1|\hat{T}, Z) = -16.061 + 2.002\ \hat{T} + 0.054\ \mathit{age} + 0.577\ \mathit{smoke} + 0.007\ \mathit{cho3}.$$

The resulting IV-estimator is equal to 2.002 with standard error 0.501 obtained by the bootstrap. Table 1.3 summarizes the results of Framingham study data with the IV method.

## 6.4 Correcting for misclassification error

Various suggestions have been made on how to correct for bias due to misclassification error (Hui and Walter, 1980; Mahajan, 2006; Lederer and Küchenhoff, 2006). All the above methods for continuous explanatory variables can roughly be adapted to the case of categorical explanatory variable subject to misclassification. In line with Section 5, correction is straightforward if one has knowledge of the sensitivity and specificity. This is for instance the case where internal or external validation data are available, or these quantities might be considered to be known through good guesses. Among many other methods, the simulation extrapolation method (SIMEX) (see Cook and Stefanski, 1994; Carroll et al., 1996) is also a tool for correcting misclassification error. The adaptation to the misclassification situation is called the MisClassification SIMEX (MC-SIMEX) approach. The R-package *mc-simex* can take into account misclassification of a categorical response or of a categorical regressor or of both. Lederer and Küchenhoff (2006) have used MC-SIMEX methods to correct for bias of the effect of smoking on chronic bronchitis when the sensitivity and specificity degree were guessed. Lewbel (2007) uses instruments to correct for misclassification in treatment effect models. Two closely related studies on misclassification in a dichotomous variable are Hui and Walter (1980) and Mahajan (2006), which use a secondary measurement or an instrument to fit a nonlinear model that includes a mismeasured binary regressor. Gustafson (2003) elaborates on how to correct misclassification error with Bayesian adjustments. In the next example, we reanalyze the effect of smoking on chronic bronchitis of some worker by the MC-SIMEX method, using the same basic idea for continuous explanatory variables (Lederer and Küchenhoff, 2006).

**Example 9**. We use data from a study on the Chronic Bronchitis and Dust concentration of the Deutsche Forschungsgemeinschaft. The data were recorded during the years 1960 and 1977 in a Munich plant (1246 workers). This contains Chronic Bronchitis reaction (1=yes, 0=no) as outcome, Smoking by self report (1=yes, 0=no) as a discrete explanatory variable, Dust concentration at work (in mg/m3) and Duration of exposure in year. We reanalyze the data set for the effect of self-reported smoking which is subject to misclassification, $W$, on Chronic Bronchitis reaction by the SIMEX method when Dust concentration at work and Duration of exposure are supposed to be error-free covariates $Z$. Specifically the aim is to estimate $\beta_1^*$, the slope coefficient of a logistic regression of Chronic Bronchitis reaction on smoking adjusted

for $Z$. We first obtain naive estimates by fitting the logistic regression model

$$\text{logit}P(Y = 1|W, Z) = -3.05 + 0.68\ \textit{smoke} + 0.09\ \textit{dust} + 0.04\ \textit{expo}.$$

Research from other studies shows that about 8% of smokers self-report them as non-smokers, that is, $\pi_{01} = P(W = 0|X = 1) = 0.08$. It is reasonable to believe that about 100% of non-smokers self-report them as non-smokers, that is, $\pi_{00} = P(W = 0|X = 0) = 1$. The misclassification matrix for this is then defined (Lederer and Küchenhoff, 2006); $\pi = \begin{bmatrix} \pi_{00} & 1-\pi_{11} \\ 1-\pi_{00} & \pi_{11} \end{bmatrix} = \begin{bmatrix} 1 & 0.08 \\ 0 & 0.92 \end{bmatrix}$
Then, we use the *mc-simex* function in R. The naive and MC-SIMEX corrected estimators and their standard error are summarized in Table 1.4. Standard error for MC-SIMEX corrected estimator is obtained by the Jackknife variance. (The data is available at www.stat.uni-muenchen.de).

Table 1.4: *MC-SIMEX approach to correct misclassification error of smoking on Chronic Bronchitis. The naive and MC-SIMEX corrected estimators and their standard error are presented.*

| Estimate | $\hat{\beta}_1$ | SE |
|---|---|---|
| Naive estimator | 0.68 | 0.173 |
| Corrected estimator | 0.88 | 0.224 |

# Appendix 1.A: Derivation of Asymptotic Bias Expressions

We derive the asymptotic bias expressions under the assumption of normality and nondifferential measurement error.

## Bias expressions when $X$ is continuous

Suppose that $\mathrm{E}(Y|X) = \beta_0^* + \beta_1^* X$, and that the data are analyzed using the working model $\mathrm{E}(Y|W) = \theta_0^* + \theta_1^* W$. By setting the expected estimating equation functions corresponding to the working model to zero:

$$\begin{aligned} 0 &= \mathrm{E}\{U(\theta^*)\} = \mathrm{E}\left\{ \begin{pmatrix} 1 \\ W \end{pmatrix} (Y - \theta_0^* - \theta_1^* W) \right\} \\ &= \mathrm{E}\left\{ \begin{pmatrix} 1 \\ W \end{pmatrix} (\mathrm{E}(Y|X,W) - \theta_0^* - \theta_1^* W) \right\}, \end{aligned}$$

we obtain the limiting values $\theta_0^*$ and $\theta_1^*$ of the ordinary least squares estimators of the intercept and slope parameters. It follows from the nondifferential measurement error assumption that

$$0 = \mathrm{E}\left\{ \begin{pmatrix} 1 \\ W \end{pmatrix} (\beta_0^* + \beta_1^* X - \theta_0^* - \theta_1^* W) \right\}.$$

The latter yields

$$\begin{aligned} \mathrm{E}(\beta_0^* + \beta_1^* X - \theta_0^* - \theta_1^* W) &= 0 \\ \mathrm{E}\{W(\beta_0^* + \beta_1^* X - \theta_0^* - \theta_1^* W)\} &= 0. \end{aligned}$$

We calculate the asymptotic bias expressions by solving the latter resulting set of equations.

1. Under the classical additive error model, $W = X + U$ with $U \perp\!\!\!\perp X$. It follows from the first equation that $\beta_0^* - \theta_0^* = \theta_1^* E(W) - \beta_1^* E(X)$. Substituting the latter into the second equation shows that,

$$\begin{aligned} \theta_1^* &= \frac{Cov(X,W)}{Var(W)}\beta_1^* \\ &= \frac{Cov(X,X+U)}{Var(W)}\beta_1^* = \frac{Var(X)}{Var(W)}\beta_1^*. \end{aligned}$$

   As a result:
   $\theta_1^* = \frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2}\beta_1^*$
   $\theta_0^* = \beta_0^* + \frac{\mu_w \sigma_u^2}{\sigma_x^2+\sigma_u^2}\beta_1^*$.

2. Under the Berkson error model, $X = W + U$, $U \perp\!\!\!\perp W$. It follows from $Cov(X,W) = Cov(W+U,W) = Var(W)$ that
   $\theta_1^* = \beta_1^*$
   $\theta_0^* = \beta_0^*$.

Suppose now that $\mathrm{E}(Y|X,Z) = \beta_0^* + \beta_1^* X + \beta_2^* Z$, and that the data are analyzed using the working model $\mathrm{E}(Y|W,Z) = \theta_0^* + \theta_1^* X + \theta_2^* Z$. Analogously, by setting the expected estimating equation functions corresponding to the working model to zero:

$$0 = \mathrm{E}\{U(\theta^*)\} = \mathrm{E}\left\{ \begin{pmatrix} 1 \\ W \\ Z \end{pmatrix} (Y - \theta_0^* - \theta_1^* W - \theta_2^* Z) \right\},$$

we obtain the limiting values $\theta_0^*$, $\theta_1^*$, and $\theta_2^*$. It follows from the nondifferential measurement error assumption that

$$\begin{aligned} 0 &= \mathrm{E}\left\{ \begin{pmatrix} 1 \\ W \\ Z \end{pmatrix} (\mathrm{E}(Y|X,W,Z) - \theta_0^* - \theta_1^* W - \theta_2^* Z) \right\} \\ &= \mathrm{E}\left\{ \begin{pmatrix} 1 \\ W \\ Z \end{pmatrix} (\beta_0^* + \beta_1^* X + \beta_2^* Z - \theta_0^* - \theta_1^* W - \theta_2^* Z) \right\}. \end{aligned}$$

This yields the following three equations:

$$\beta_0^* - \theta_0^* + \beta_1^* E(X) - \theta_1^* E(W) + (\beta_2^* - \theta_2^*)E(Z) = 0 \tag{1.18}$$
$$(\beta_0^* - \theta_0^*)E(W) + \beta_1^* E(WX) - \theta_1^* E(W^2) + (\beta_2^* - \theta_2^*)\mathrm{E}(WZ) = 0$$
$$(\beta_0^* - \theta_0^*)E(Z) + \beta_1^* E(ZX) - \theta_1^* E(WZ) + (\beta_2^* - \theta_2^*)\mathrm{E}(Z^2) = 0. \tag{1.19}$$

It follows from the first equation that $\theta_0^* = \beta_0^* + \beta_1^* \mu_x - \theta_1^* \mu_w + (\beta_2^* - \theta_2^*)\mu_z$. By plugging into the second equation, we obtain $(\beta_2^* - \theta_2^*)Cov(W, Z) = \theta_1^* \sigma_w^2 - \beta_1^* Cov(X, W)$. Next plugging both into the third equation, we obtain

$$\theta_1^* = \frac{\sigma_z^2 Cov(X, W) - Cov(X, Z)Cov(W, Z)}{\sigma_z^2 Var(W) - Cov^2(W, Z)}\beta_1^*.$$

Under the following assumptions,

$$\begin{aligned} Z &\sim N(\mu_z, \sigma^2) \\ U &\sim N(0, \sigma_u^2) \\ W|Z &\sim N(\eta_0^* + \eta_1^* Z, \sigma_{w|z}^2) \\ X|Z &\sim N(\eta_0^* + \eta_1^* Z, \sigma_{x|z}^2) \\ E(W|Z) &= E(X|Z) \end{aligned}$$

together with the classical additive error model, $W = U + X$ with $U \perp\!\!\!\perp (X, Z)$, we obtain

$$\begin{aligned} E(XW|Z) &= E(X^2|Z) \\ Var(W|Z) &= Var(X + U|Z) = Var(X|Z) + Var(U) \\ Cov(W, U|Z) &= E(WU|Z) = E\{(X + U)U|Z\} = \sigma_u^2 \\ Cov(W, Z) &= E\{ZE(W|Z)\} - E(Z)E\{E(W|Z)\} = \eta_1^* \sigma_z^2 = Cov(X, Z). \end{aligned}$$

After some algebraic operations, we find

$$\begin{aligned} \theta_0^* &= \beta_0^* + \frac{\sigma_u^2}{\sigma_{x|z}^2 + \sigma_u^2}\beta_1^* \eta_0^* \\ \theta_1^* &= \frac{\sigma_{x|z}^2}{\sigma_{x|z}^2 + \sigma_u^2}\beta_1^* \\ \theta_2^* &= \beta_2^* + \frac{\sigma_u^2}{\sigma_{x|z}^2 + \sigma_u^2}\beta_1^* \eta_1^*. \end{aligned}$$

## Bias expressions when $X$ is discrete

We now derive similar asymptotic bias expressions due to misclassification error on the dichotomous explanatory variable in the linear model $\mathrm{E}(Y|X,Z) = \beta_0^* + \beta_1^* X + \beta_2^* Z$, under the nondifferential measurement error. As stated, we suppose the misclassification probabilities are not related to the error-free covariate $Z$. That is, $P(W=1|X=1,Z) = \pi_{1|1}$ and $P(W=0|X=0,Z) = \pi_{0|0}$. Let $P(W=1|X,Z) = a + bX$, where $b = \pi_{1|1} + \pi_{0|0} - 1$ and $a = 1 - \pi_{0|0}$. Similarly, by setting the expected estimating equation functions corresponding to the working model to zero:

$$0 \quad = \quad \mathrm{E}\left\{U(\theta^*)\right\} = \mathrm{E}\left\{ \begin{pmatrix} 1 \\ W \\ Z \end{pmatrix} (Y - \theta_0^* - \theta_1^* W - \theta_2^* Z) \right\},$$

we obtain

$$\theta_1^* = b \left[ \frac{\mu_x(1-\mu_x)(1-\rho_{xz}^2)}{\mu_w(1-\mu_w) - \mu_x(1-\mu_x)b^2\rho_{xz}^2} \right] \beta_1^*$$

$$\theta_2^* = \beta_2^* + \beta_1^* \rho_{xz} \frac{\sqrt{\mu_x(1-\mu_x)}}{\sigma_z} \left[ 1 - b\frac{\theta_1^*}{\beta_1^*} \right]$$

upon noting that, $Var(X) = \mu_x(1-\mu_x)$, $Var(W) = \mu_w(1-\mu_w)$, and

$$\begin{aligned} \mu_w &= P(W=1) = a + b\mu_x \\ \mathrm{E}(XW) &= \mathrm{E}\{X\mathrm{E}(W|X,Z)\} = (a+b)\mu_x \\ Cov(X,W) &= b\mu_x(1-\mu_x) \\ \mathrm{E}(ZW) &= \mathrm{E}\{Z\mathrm{E}(W|X,Z)\} = a\mu_z + bE(ZX) \\ Cov(W,Z) &= bCov(X,Z) = b\rho\sigma_x\sigma_z \end{aligned}$$

where $\rho_{xz}$ is the correlation between $X$ and $Z$.

# Chapter 2

# Introduction to Causal Inference

## Summary

This chapter gives an introduction to the statistical framework of counterfactual outcomes for causal inference as it has been developed over the past decades. We define causal effects and causal models formally in this setting, and state the assumptions which allow to identify causal effects. We explain the problem of confounding and discuss methods to control for measured as well as unmeasured confounding. We introduce the special clan of problems of noncompliance in randomized controlled trials and expand on methods to adjust for it. We describe the instrumental variables approach to allow for inference on the causal effect of exposure on outcome in the presence of noncompliance in randomized controlled trials, and for unmeasured confounder adjustment in observational studies. We continue to investigate the problem of measurement error in exposure and important practical implications under linear structural mean models. For detailed exposition and technical aspects of the developed methods, we refer to the original literature. Basic definitions and notions are described here, others will be introduced over the next chapters, as necessary.

# 1 Introduction

Causal inference deals with cause-effect relationships between interventions (exposures) and outcomes (responses) in many fields of study including biostatistics, epidemiology and econometrics. The goal of causal inference is typically to estimate the causal effects of interventions or exposures on outcomes of interest from empirical data through design and analysis. For instance in biostatistics, one might be interested to estimate the causal effect of a specific drug or other medical intervention on a primary health or risk outcome. In epidemiology, for instance, researchers are concerned with the evaluation of the effect of AZT on HIV viral load or CD4 in a specific population which is HIV infected. In econometrics, one may wish to evaluate the effect of educational training on income. In settings as this, statisticians have known that the detection of any causal effect of an intervention or exposure on an outcome must rely on knowledge about how experiments were performed or how more generally the data were generated. In the ideal experiment each subject receives all interventions or exposures under investigation during a given time interval or at one specific time point, and the different results under each exposure can be compared. In practice, this is unfortunately typically impossible and other approaches must be developed.

Detection and quantification of causal effects forms an ideal evidence basis for policy decisions and interventions. In several major case studies however, standard statistical methods have failed. A new formal statistical development of causal inference was therefore more than welcome when it was first introduced in the 1970's (Rubin, 1974, 1978; Robins, 1986; Holland, 1988; Pearl, 1988). Results of this era of innovative theory are now disseminating to applied fields, frequently resulting in important advances. There are at least two explicit sets of mathematical language for approaching causal inference. One evolves around the counterfactual outcome model of Neyman-Rubin, as well as Robins's extensions (Neyman, 1923; Rubin, 1974; Robins, 1987). More recently, a second set involves a combination of structural equations together with a graphical representation, which takes the form of a causal directed acyclic graph (DAG) to communicate causal relationships (Pearl, 1995, 2000). The counterfactual outcome model is more explicit for obtaining estimators and estimands. It is an effective mathematical formalism allowing for parameterization and to clearly state assumptions which might be needed to identify causal effects. The causal directed acyclic graphs are also effective tools to check assumptions on the observed relation between exposure and outcome, and to recognize the likely factors that may bias the estimated causal effect of exposure on outcome. These are particularly useful to understand relations in the complex setting of time-varying exposures and outcomes in longidudinal studies.

# 2 Causal effect and counterfactual outcomes

At the individual level, a causal effect could be defined as a contrast between the possible results from two or more alternative exposures or interventions with only one of the results actually observed. One thus considers whether and how much the outcome would have been different had the exposure been different, all other conditions being the same. For example, suppose that a child who lives near a chemical factory contracts a rare cancer. Suppose that, we seek to establish whether or not a chemical spill adjacent to the child's property was the cause of his/her particular cancer (Spirtes et al., 2000). By saying that the chemical spill caused the disease on this individual level, we mean that the cancer would not have occurred had, contrary to fact, the spill not happened.

In order to avoid unnecessary complications and without loss of generality, we will conduct our first discussion in a very simple setting, for example considering $i = 1, ..., n$, subjects which are exposed to a hypothetical dichotomous exposure $X_i$ with two possible levels ($X_i = 1$ if exposed, 0 if unexposed). Suppose that half of these subjects have been exposed to 1, an experimental treatment, and the other half have been exposed to 0, no experimental treatment which may be placebo or a standard treatment. Suppose that the experimental treatment $X_i$ is a single dose of vitamin $A$ versus the control treatment (placebo), and the subject $i$ is a particular child. Let $x \in \{0, 1\}$ be an index indicating the possible counterfactual receipt of exposure $X_i$, we then define $Y_{ix}$ as the outcome that child $i$ would have had if he/she received treatment $X_i = x$ for $x = 0, 1$. Once the child received the vitamin $A$, we assume that $Y_{i1}$ for that child equals his/her observed outcome $Y_i$, and $Y_{i0}$ is unobserved for this child. The variables $Y_{i1}$ and $Y_{i0}$ are called counterfactual or potential outcomes because one of them (namely $Y_{i0}$ in this case) describes the subject $i$'s outcome value that would have been observed under a potential exposure value which (i.e., $X_i \neq 0$) did not actually occur (Rubin, 1974, 1978; Holland, 1988; Robins, 1986; Vansteelandt and Goetghebeur, 2003). Note that, $Y_{i0}$ is often simply called treatment-free outcome (identical to $Y_i$ in those receiving no treatment, mainly those in the control group, but a counterfactual in the treated).

Throughout this chapter, we shall write $Y_i$ for the measured outcome, $X_i$ for the observed exposure or treatment, and $Y_{ix}$ for the counterfactual or potential outcome where $X_i$ is set to a possible exposure level $x$. We use $Z_i$ to encode for the baseline covariates for each subject $i$. In the general case, however, outcome and exposure are not limited to be dichotomous as considered in the above hypothetical example. If the outcome is continuous, for instance the amount of blood pressure reduction under a possible dose of drug, then there is a correspond set of continuous counterfactual outcomes indexed by the drug dosage. We are now ready to state our definitions more

formally.

**Individual causal effect**. The individual causal effect for a given subject $i$, for instance, exposed to treatment $X_i = 1$ versus placebo $X_i = 0$ (unexposed), all other things being equal, can be defined as the difference between counterfactual outcomes, corresponding to the different possible treatment levels,

$$Y_{i1} - Y_{i0}.$$

If the outcome variable is non zero, we may instead define the causal effect as the ratio

$$Y_{i1}/Y_{i0},$$

or when it is strictly positive as an outcome difference on the log scale

$$\log Y_{i1} - \log Y_{i0}.$$

In words, an individual causal effect is a contrast between counterfactual outcomes of a single subject $i$ under different possible treatment levels. It is generally assumed that $Y_{i1} - Y_{i0}$ need not remain the same from one subject to another and, in particular, one might be interested in investigating how this contrast might be associated with treatment actually received. In general, there is not enough information to identify an individual causal effect because counterfactual outcomes of an individual cannot be observed under both the treatment and the control conditions simultaneously. That is, we cannot observe both counterfactual outcomes $Y_{i1}$ and $Y_{i0}$ for the subject $i$, but we do observe $Y_i = (1 - X_i)Y_{i0} + X_i Y_{i1}$. This assumption is commonly referred to as the consistency assumption, which states that the observed outcome coincides with the corresponding counterfactual outcome.

Since one of the counterfactual outcomes is always missing we cannot estimate the individual causal effect of exposure, $Y_{i1} - Y_{i0}$, structurally for any of subject $i$. As this definition reveals, causal inference can be seen as a problem of missing data. Therefore, one aims to estimate the causal effect for a population instead, by comparing the distribution of $Y_i$ for subjects with different levels of exposure $X_i$ which may require weaker assumptions.

**Population causal effect**. At the population level, a causal effect is defined by comparison of features of the distribution of counterfactuals such as the mean, specific probabilities, median, etc.. In particular, for a dichotomous exposure we will see that a population causal effect is for instance defined by a difference of expected counterfactual outcomes

$$\mathrm{E}(Y_{i1} - Y_{i0}),$$

or for the positive outcome variable by

$$\log \mathrm{E}(Y_{i1}) - \log \mathrm{E}(Y_{i0}).$$

The estimand is typically the average causal effect (ACE) of exposure when the outcome is continuous, or could be the causal risk ratio (CRR), or the causal odds ratio (COR) when the outcome is dichotomous. These quantities are explained in the next section. Note that the above definitions do not involve the joint distribution under the potential exposures, but merely the marginal distribution under each possible exposure or treatment level.
While one is often investigated in the average difference between treated and untreated outcomes across all subjects in a population, one may alternatively wish to estimate a conditional average causal effect (CACE), for instance conditional on observed treatment

$$\mathrm{E}(Y_{i1} - Y_{i0}|X_i). \tag{2.1}$$

In some situations, one may also be interested in the average causal effect within a subset of the population determined by observed covariates $Z_i$, e.g.

$$\mathrm{E}(Y_{i1} - Y_{i0}|X_i, Z_i).$$

For example, one might want to know the average causal effect of male alcohol consumption on oesophageal cancer risk. Thus, the causal effect can easily be adapted by conditioning on covariates provided these are prior to exposure (pre-treatment).

## 2.1 Causal parameters and causal assumptions

When the distributions of two counterfactual or potential outcomes differ under the possible levels of different observed exposure $X$, we say that the exposure $X$ has a causal, causative or preventive effect on the population. The corresponding causal effects of interest are generally described by parameters under a statistical (causal) model for the counterfactual outcomes. For instance, when $Y$ is a continuous outcome, ACE is a natural choice of causal parameter when the effect of exposure is suspected to be linear on outcome. Suppose that, one wishes to estimate the effect of college graduation, which is represented by the dichotomous variable $X_i$, on some outcome $Y_i$ of interest, say earnings (Rubin, 1997; Abadie, 2003). In this example, $Y_{i1}$ represents the potential earnings as a college graduate while $Y_{i0}$ represents the potential earnings as a non-graduate. One may consider the following model to infer the average causal effect on the treated

$$\mathrm{E}(Y_{i1} - Y_{i0}|X_i = 1) = \psi^* \tag{2.2}$$

where $\psi^*$ is the causal parameter which is unknown and can be estimated. We only observe $Y_i = X_i Y_{i1} + (1 - X_i)Y_{i0}$ under the consistency assumption and comparisons of earnings for the graduated and non-graduated do not usually yield an unbiased causal effect of college. Indeed, under the consistency assumption we have

$$\begin{aligned} \mathrm{E}(Y_i|X_i = 1) - \mathrm{E}(Y_i|X_i = 0) &= \mathrm{E}(Y_{i1}|X_i = 1) - \mathrm{E}(Y_{i0}|X_i = 0) \\ &= \mathrm{E}(Y_{i1} - Y_{i0}|X_i = 1) + \mathrm{E}(Y_{i0}|X_i = 1) \\ &\quad -\mathrm{E}(Y_{i0}|X_i = 0) \\ &= \psi^* + \mathrm{E}(Y_{i0}|X_i = 1) - \mathrm{E}(Y_{i0}|X_i = 0). \end{aligned}$$

The term $\mathrm{E}(Y_{i0}|X_i = 1) - \mathrm{E}(Y_{i0}|X_i = 0)$ represents the bias caused by endogenous selection in the treatment. That is, in a selective sample common unobservables may affect both the outcome and the probability of exposure in unknown ways (Abadie, 2003; Lewbel, 2007). In general, this bias is different from zero. If we assume that $Y_{ix} \perp\!\!\!\perp X_i$, that is, observed treatment and counterfactual outcomes are independent for all $x$, then

$$\mathrm{E}(Y_{ix}) = \mathrm{E}(Y_{ix}|X_i = x) = \mathrm{E}(Y_i|X_i = x).$$

This implies that $\mathrm{E}(Y_{i0}|X_i = 1) - \mathrm{E}(Y_{i0}|X_i = 0) = 0$ and then $\psi^*$ can be estimated. However, independence between treatment and counterfactual outcomes may not be satisfied in most applications where selection for treatment is not random. In some cases, this assumption is plausible once we condition on a vector of observed covariates $Z_i$. This situation is called selection on observables (Heckman and Robb, 1985). In model (2.2), when the exposure has no causal effect, we say that the sharp causal null hypothesis, $\psi^* = 0$, is true.

Similarly, when outcome is dichotomous, we define the probability $P(Y_{ix} = 1)$ as the proportion of subjects that would have developed the outcome had all subjects in the population of interest received exposure value $x$. It is then more common to infer the causal risk ratio (CRR) defined as

$$P(Y_{i1} = 1)/P(Y_{i0} = 1),$$

or the causal odds ratio (COR) given by

$$\frac{P(Y_{i1} = 1)/P(Y_{i1} = 0)}{P(Y_{i0} = 1)/P(Y_{i0} = 0)}.$$

The risk ratio and odds ratio are effect measures which can be used to quantify the strength of the causal effect when it exists. They measure the same causal effect on different scales. An other causal parameter is called the intention-to-treat parameter.

This parameter represents the average causal effect obtained by assigning everyone in the study population to treatment, rather than the control regimen in the context of a randomized controlled trial (explained further in next section).
To be able to estimate the causal effect, we need to involve assumptions that are sometimes mathematically untestable based on the data. Examples of such assumptions include the assumption of no unmeasured confounders in observational studies (e.g., Rubin, 1978; Robins et al., 1992) and the assumption that the counterfactual value $Y_{ix}$ for subject $i$ does not depend upon the treatments received by other subjects $j \neq i$. This is commonly referred to as the stable unit treatment value assumption (SUTVA) (e.g., Rubin, 1990; Angrist et al., 1996). We will explain these and other assumptions throughout the text as necessary.

## Causation versus association

We say $X$ and $Y$ are statistically independent if knowing the value of $X$ does not provide any information about the value of $Y$, and statistically associated or dependent if knowing the value of $X$ provides some information about the likely value of $Y$, even if this information is very limited and amounts to a modest change in the probability distribution of $Y$. Scientific studies show an association between an independent variable and dependent variable do not necessarily imply that the independent variable causes the dependent variable. For example, if dementia is more prevalent among people with limited education, then knowing someone's education provides information, albeit not certainty, on the probability that the person has dementia (Glymour et al., 2008). Thus we say education and dementia are statistically associated, which is quite different from saying that (lack of) education causes dementia. Pearl (2001) says "associations characterize static conditions, while causal analysis deals with changing conditions". To show that association is not generally causation, we will adapt an example from Hernán (2004) for illustration. In this example, the treatment $X = 1$ when a patient receives a heart transplant and 0 otherwise, and the outcome $Y = 1$ for death and 0 otherwise. The data are summarized in Table 2.1. The probability $P(Y_x = 1)$ is defined as the proportion of subjects that would have died had all subjects in the population of interest received exposure value $x$, hence the risk of $Y_x$. As Table 2.1 shows, the exposure has no causal effect in the population because $P(Y_1 = 1) = P(Y_0 = 1) = 0.50$, that is, $P(Y_1 = 1) - P(Y_0 = 1) = 0$. In reality however, only one of both potential outcomes is observed. The question is then how to infer the causal effect of interest, despite the missing data on either $Y_0$ or $Y_1$. When measuring the association between exposure $X$ and outcome $Y$, we calculate the proportion of subjects that underwent treatment $X = 1$ (heart transplant) and compare it with the proportion of subjects that underwent treatment $X = 0$

(no heart transplant). Formally, one thus examines whether $P(Y = 1|X = 1)$ equals $P(Y = 1|X = 0)$. We can easily find that $P(Y = 1|X = 1) = 7/13$ differs from $P(Y = 1|X = 0) = 3/7$. It follows that there is an association between exposure

Table 2.1: *Counterfactual outcomes of subjects in a study with dichotomous exposure X (subtable left); unobserved counterfactual outcomes, dichotomous exposure X and observed outcome Y (subtable right). Sign ? indicates missingness of the counterfactual outcomes (Hernán, 2004).*

| Subject | $Y_0$ | $Y_1$ | $X$ | $Y$ | $Y_0$ | $Y_1$ |
|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 0 | 0 | ? |
| 2 | 1 | 0 | 0 | 1 | 1 | ? |
| 3 | 0 | 0 | 0 | 0 | 0 | ? |
| 4 | 0 | 0 | 0 | 0 | 0 | ? |
| 5 | 0 | 0 | 1 | 0 | ? | 0 |
| 6 | 1 | 0 | 1 | 0 | ? | 0 |
| 7 | 0 | 0 | 1 | 0 | ? | 0 |
| 8 | 0 | 1 | 1 | 1 | ? | 1 |
| 9 | 1 | 1 | 0 | 1 | 1 | ? |
| 10 | 1 | 0 | 0 | 1 | 1 | ? |
| 11 | 0 | 1 | 0 | 0 | 0 | ? |
| 12 | 1 | 1 | 1 | 1 | ? | 1 |
| 13 | 1 | 1 | 1 | 1 | ? | 1 |
| 14 | 0 | 1 | 1 | 1 | ? | 1 |
| 15 | 0 | 1 | 1 | 1 | ? | 1 |
| 16 | 0 | 1 | 1 | 1 | ? | 1 |
| 17 | 1 | 1 | 1 | 1 | ? | 1 |
| 18 | 1 | 0 | 1 | 0 | ? | 0 |
| 19 | 1 | 0 | 1 | 0 | ? | 0 |
| 20 | 1 | 0 | 1 | 0 | ? | 0 |

$X$ and outcome $Y$. If a risk difference is chosen to report the association, it equals $P(Y = 1|X = 1) - P(Y = 1|X = 0) = 7/13 - 3/7 = 0.11$ which differs from the causal effect of zero. For this data, $\frac{P(Y=1|X=1)}{P(Y=1|X=0)} = 1.26$ and $\frac{P(Y=1|X=1)/P(Y=0|X=1)}{P(Y=1|X=0)/P(Y=0|X=0)} = 1.56$. These are the association risk ratio (ARR) and the association odds ratio (AOR), respectively, which may differ from the CRR and COR. Unlike association measures, causal effect measures cannot be directly computed because of missing data (Table

2.1, subtable right). We will show (in the next section), under the randomization assumption $Y_x \perp\!\!\!\perp X$, that $P(Y = 1|X = x) = P(Y_x = 1)$. Some authors often use the "do" operator by $P(Y|do(X = x))$, to distinguish between conditioning on an exposure in $X$ and the usual conditioning on observing $X$, $P(Y|X = x)$ (Pearl 1995, 2001; Didelez et al., 2007, 2008); we however use an index for the former.

## Experimental studies

A study, where for instance a dichotomous exposure $X$ ($X = 1$ exposed, $X = 0$ unexposed) is randomly assigned to $n$ subjects (using some mechanism that assured each subject was equally likely to be exposed and unexposed) is called a randomized experiment or more simply an experimental study. In randomized experiments, the exposed and unexposed groups are comparable a priori (under the assumption of perfect compliance which is explained in next section). This implies that if subjects were randomly assigned to group $X = 1$ and $X = 0$, the proportion of subjects with counterfactual outcome $Y_1 = 1$ among the exposed will be the same in group $X = 1$ as in group $X = 0$. Thus, which particular group $x$ got the exposure is irrelevant for the value of $P(Y_1 = 1|X = x)$ for $x = 0, 1$. Formally, we say that both groups are exchangeable. Due to exchangeability in ideal, randomized experiments, association is causation (Hernán, 2004; Hernán and Robins, 2006). In particular, association measures can be interpreted as effect measures of the assignment of the randomized treatments. Under noncompliance however, the assigned treatment does not necessarily coincide with treatment actually received. In any case, under randomization $Y_x \perp\!\!\!\perp X$; but only with prefect compliance this translates into $P(Y_x = 1) = P(Y = 1|X = x)$. In Table 2.1, for instance for $x = 1$, $P(Y_1 = 1) = P(Y = 1|X = 1) = 7/13$. In the next section, we will elaborate on noncompliance in the randomized experiments.

## Observational studies

Investigators often use observational studies when controlled experiments are infeasible or unethical. For instance, in the epidemiological setting randomized controlled trials are unlikely to be carried out to evaluate the effect of complex nutritional regimes or it is unethical to asses the effect of smoking and alcohol consumption by randomly allocating individuals to these exposures. Instead, in observational studies which are also called nonrandomized studies or sometimes quasi-experiments, the exposed and unexposed groups are generally not comparable and not exchangeable. One thus cannot estimate causal effects merely by calculating associations without making additional assumptions (Pearl, 2000; Hernán, 2004; Hernán and Robins, 2006). A major challenge in conducting observational studies is to draw inferences that are

acceptably free from overt biases, as well as to assess the influence of potential hidden biases. One basic solution to this problem is to adjust for pertinent confounding (Robins et al., 2000; Henneman et al., 2002; Greenland, 2000). In the next chapter, we investigate in more detail the problem of unmeasured confounders in observational studies. That is, adjusting the effect of variables which might be confounded in the effect of exposure on the outcome of interest.
In general, determining an appropriate experiment to use to detect causal parameter depends upon the circumstance of exposures and the outcome of interest. The causal effect estimate of an exposure might be more robust when the exposure of interest can be randomly assigned (Glymour et al., 2008). The strength of this design arises because randomization essentially eliminates common prior causes (confounding) as an explanation for a statistical association. For example, many observational studies have found that postmenopausal hormone therapy (HT) reduced the risk for coronary heart disease (CHD). A randomized study of HT reported an increased risk for recurrent CHD among HT users, the discrepancy was interpreted as evidence that experimentation trumps observation (Nananda et al., 2003). Consider also, for example, the effect of antioxidant vitamin intake on the risks of cancer, cardiovascular disease, and mortality. Observational studies have shown protective effects against these outcomes (Khaw et al., 2004), while in contrast, randomized trials have shown no effect. It has been suggested that the disparity in results is likely due to confounding by behavioral and social factors acting across the life course. For example, factors related to childhood social class may be important confounders of the association between antioxidant vitamin intake and disease outcome.

## 2.2 Causal directed acyclic graphs (DAGs)

Directed acyclic graphs (DAGs) are a useful graphical tool in causal inference that can help to detect the causal effect of an exposure on an outcome in the presence of other variables. They in fact graphically encode all conceived causal influences between all observed variables in an experiment of interest. In particular, the DAG together with structural equation models allows to check the causal assumptions that one relies on to identify the causal effect of exposure on outcome. They can also reveal which measured and unmeasured factors are critical confounders of the actual exposure effect on outcome. In this section, we briefly pay attention to the causal directed acyclic graph (DAG) as used by Pearl (1995, 2000, 2001) (Greenland et al., 1999; Robins et al., 2000).
A DAG is a set of vertices (or nodes) and a set of edges (arrows) that connect pairs of these vertices. The vertices corresponds to variables and the edges will denote a certain relationship that holds between pairs of variables. For example, we might

have a set of three vertices: V1, V2, V3, and a set of two edges among these vertices: $(V1 \rightarrow V2, V2 \rightarrow V3)$. A DAG represents a picture, or path diagram. For example, this DAG looks like: $V1 \rightarrow V2 \rightarrow V3$. In a DAG, each edge represents a direct causal relationship, no spurious association. The DAG in fact encodes the possible existence of (direct or indirect) causal influence between all relevant variables, and the absence of a (direct or indirect) causal influences between them. Besides representing causal relations, causal DAGs also encode the causal determination of statistical associations.

- If exposure $X$ causes outcome $Y$, then $X$ and $Y$ will generally be associated. Figure 1 represents a follow-up study of HIV-infected patients where $X$ is the dose of the treatment, say AZT, and $Y$ indicates for example the presence of detectable HIV RNA; that is, $Y = 1$ if HIV RNA is detectable in the blood and is 0 otherwise.

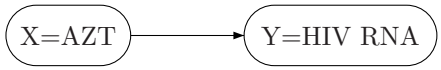



Figure 1: *Causal acyclic graph indicating that AZT causes HIV RNA.*

- If exposure $X$ and outcome $Y$ share a common cause, then $X$ and $Y$ will generally be associated, even if neither is the cause of the other. Figure 2 shows that $Z$ =(CD4, age, etc.) is a common cause of AZT and HIV RNA.

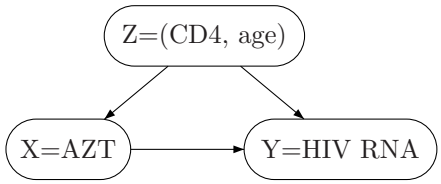



Figure 2: *Causal acyclic graph that Z (such as CD4 lymphocyte, age, etc.) is a common cause of AZT and HIV RNA.*

- A direct cause of variable $V$ on the DAG is called a *parent* of $V$ and $V$ is called a *parent's child*. In Figures 1 and 2, AZT is a parent of HIV RNA and HIV RNA is a child of AZT.

Following Pearl (1995, 2001), a causal DAG is a set of nodes $V = (V1, ..., Vn)$, corresponding to variables, and directed edges amongst nodes such that no variable $V_i$ can cause itself (acyclic). A path is a sequence of nodes connected by edges regardless of arrowhead direction; a directed path is a path which follows the edges in the direction indicated by the graph's arrows. Further definitions,

- The *ancestors* of $V$ are those variables ancestors with $V$ as a *descendant* but are a direct cause only of its children (where direct is always relative to the other variables on the DAG). Thus, $V$ is caused by all its ancestors, but only its parents are direct causes.
- A path collides at a variable $V$ if the path enters and exits $V$ through arrowheads, in which case $V$ is called a *collider* : $\rightarrow V \leftarrow$.
- There are different types of paths between a variable $V_1$ and a variable $V_2$:
  - A directed path from $V_1$ to $V_2$, as in Figure 1
  - A backdoor path from $V_1$ to $V_2$: a path whose first edge is an arrow pointing to $V_1$ and whose last edge is an arrow pointing to $V_2$.
  - A blocked path between $V_1$ and $V_2$ : a path that has one or more colliders; otherwise it is unblocked or open. Thus, a back-door path and a directed path are open paths.

To explain the above terminologies consider Figure 3. In this Figure, $X$ is not a parent of $Y$ and $Y$ is not a child of $X$ since there is no arrow from $X$ to $Y$. However, there is a directed path from $X$ to $Y$ through $E$, so $X$ is an ancestor of $Y$ and $Y$ is a descendant of $X$. There is a back-door path from $X$ to $Y$ through $Z$ and a blocked path between $X$ and $Y$ because $L$ is a collider on the path from $X$ to $Y$ going through $E$ and $L$. Finally, there are 2 open paths between $X$ and $Y$ ; one through $Z$ (the back-door path) and one through $E$ (the directed path). In Figure 1, the AZT to HIV RNA arrow indicates a direct effect of AZT on HIV RNA, and shows no confounders of the effect of AZT on HIV RNA. Figure 2 shows a mediated effect of CD4 on HIV RNA in addition to the direct effect of AZT.

**Formal definition**. Let $\bar{V}_{m-1} = (V_1, ..., V_{m-1})$ be the vector of 'ancestors' of $V_m$, $PA_m$ be the vector of 'parents' of $V_m$, and $V_m(\bar{v}_{m-1})$ is a function of $\bar{v}_{m-1}$ alone through the values of the parents. A DAG satisfies the following conditions,

- All one-step counterfactuals $V_m(\bar{v}_{m-1})$ exist.
  - each directed edge represents a stable and autonomous physical relationship. That is, it is conceivable to change such relationship without changing the others.

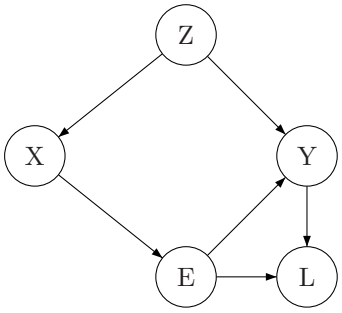



Figure 3: *An example of a DAG.*

  - there is reasonable agreement as to the hypothetical intervention which sets $X$ equal to $x$.

- $V_m(\bar{v}_{m-1}) = V_m(pa_m)$ is a function of $\bar{v}_{m-1}$ alone through the values of the parents. Once all parents are manipulated, no further variable can have an influence. For example, $X \to Z \to Y$ expresses that $Y_{xz} = Y_{x^*z} = Y_z$. This is called the restriction assumption that says $X$ has no direct effect on $Y$ except through $Z$.

- Observed variables and counterfactuals are obtained recursively from $V_m(\bar{v}_{m-1})$. For example, $X \to Z \to Y$ expresses that $Y_x = Y_{xz(x)} = Y_{z(x)}$, and $Z = Z(x)$ and $Y = Y_x = Y_{xz(x)} = Y_{z(x)}$. This is the consistency assumption that links potential outcomes to the observed data.

- All common causes of any two variables are included as variable on the DAG. This is called the no omitted confounders assumption.

$$\{V_{m+1}(\bar{v}_m), ..., V_n(\bar{v}_{n-1})\} \perp\!\!\!\perp V_m | \bar{V}_{m-1} = \bar{v}_{m-1}$$

  where $\bar{v}_{m-1}$ is subvector of $\bar{v}_k$, $k \geq m$.

As stated, the central question in the analysis of the causal effect is: can the controlled (post-intervention or post-treatment), $f(Y_x)$, be estimated from the pre-intervention

distribution $f_{Y|X,Z}(y|x,z)$? "A fundamental theorem in causal analysis states that such identification would be feasible whenever the causal model is Markovian, that is, the graph is acyclic (containing no directed cycles) and all error terms are jointly independent"(Pearl, 2001).
**Theorem 1**.(*The causal Markov conditions*)
The probability density P of the observables is Markov relative to the DAG if,

$$P(V_1, ..., V_n) = \prod_{m=1}^{n} P(V_m|PA_m)$$

or equivalently, for $m = 1, ..., n, \quad P(V_m|\bar{V}_{m-1}) = P(V_m|PA_m)$.
For example for the variables in Figure 2, the joint density can be factorized

$$f_{Y,X,Z}(y, x, z) = f_{Y|X,Z}(y|x, z)f_{X|Z}(x|z)f_Z(z).$$

**Intervention density**

As stated, our initial goal was to identify the intervention density

$$P(V_1(x), ..., V_n(x))$$

where $X \subset V$.
**Corollary**. (*Truncated factorization*)
For any Markovian model, the distribution generated by an intervention $x$ on a set of $X$ is given by the truncated factorization

$$P(V_1(x), ..., V_n(x)) = \prod_{m|V_i \neq X}^{n} P(V_m|PA_m)I(X = x).$$

It follows that the intervention density of $V(x)$ is identified from the observed data under the causal graph. Under the DAG of Figure 2, we can write the intervention density of $V(x)$ as,

$$f(Y_x, X_x, Z_x) = f(Y|X = x, Z)f(Z)I(X = x)$$

then

$$f(Y_x) = \int f(y|X = x, Z = z)f_Z(z)dz.$$

In particular, we can identify the average causal effect of $X$ on $Y$ (relative risk or odds ratio causal effect in case of dichotomous outcome $Y$)

$$\begin{aligned} \mathrm{E}(Y_1 - Y_0) &= \int\int yf(y|X=1,Z=z)f(z)dydz - \int\int yf(y|X=0,Z=z)f(z)dydz \\ &= \int \{\mathrm{E}(Y|X=1,Z=z) - \mathrm{E}(Y|X=0,Z=z)\}f(z)dz. \end{aligned}$$

We are now able to derive any intervention density under a given causal DAG. While observed outcome is replaced by counterfactual outcome $Y_x$, the direct arrow from AZT to HIV RNA is removed to make $Y_x$ a non descendant of $X$. Figure 4 shows this. On the causal DAGs, the independence can be verified by using *d-separation*.

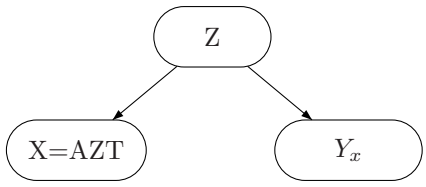



Figure 4: *Causal acyclic graph that $Z$ is measured confounder (CD4 lymphocyte, age, etc.) and replace observed outcome (HIV RNA), $Y$, by potential outcome $Y_x$.*

D-separation is a relation between three disjoint sets of vertices in a directed graph G. "To test whether $X$ is independent of $Y$ given $Z$ in any distribution compatible with G, we need to test whether the nodes corresponding to variables $Z$ block all path from nodes in $X$ to nodes in $Y$" (Pearl, 2000). Blocking is to be interpreted as stopping the flow of information (or dependency) between the variables that are connected by such paths. "The intuition behind d-separation is simple and can best be recognized if we attribute causal meaning to arrows in the graph" (Pearl, 2000). A set $C$ is said to d-separate $A$ from $B$ if and only if $C$ blocks every path from a node in $A$ to a node in $B$. In Figure 4, $Y_x$ and $X$ are d-separated by $Z$, because all paths connecting $X$ and $Y_x$ are blocked by $Z$. Then, $X$ and $Y_x$ are conditionally independent for all $x$, given $Z$; that is, $Y_x \perp\!\!\!\perp X|Z$ for all $x$. This is referred to as no unmeasured confounders. As a result, if $Y_x$ and $X$ are not associated within levels of $Z$, then the conditional association between $X$ and $Y$, given $Z$, reflects the causal effect of $X$ on $Y$ within levels of $Z$. Otherwise, the conditional association between $X$ and $Y$, given $Z$, must be spurious.

# 3 The problem of confounding in observational studies

In causal inference, as explained, one searches for the causes of an outcome, based on associations with various variables that are measured in the study. In addition to the exposure that the study is investigating, there may be other factors that are associated with the exposure and independently affect the outcome. For instance, in a study on the health effects of smoking, smokers groups may be older than non-smokers, and more of them may be male. That would confound the relationship between smoking and heart disease, since even non-smoking older males are at high risk of heart disease. If the prevalence of these factors differs between the smokers and non-smokers being compared, this will yield a biased estimate of the causal effect of exposure on the outcome of interest. These factors are called confounding factors or variables. In most standard literature a confounder is defined as a variable associated with exposure, and associated with outcome conditional on exposure and not on the causal pathway between exposure and outcome. In Figure 2, $Z$ =(CD4, age and etc.) is confounder for the association between AZT and HIV RNA. It is important to control for confounding by measuring the known confounders and including them as covariates in the data analysis. However, one major problem is that confounding variables are not always measured or even imaginable. VanderWeele, Hernán and Robins (2008) investigate the sign of the bias that arises when control for confounding is inadequate. Further, they show that if only one unmeasured confounding variable is present in the data setting, it is relatively easy to draw a conclusion about the direction of the bias. If there is more than one unmeasured confounding variable, this becomes much more complex. In view of this, there are various methods which have been proposed for adjusting for measured confounders. These different methods may produce different estimates for the causal effect of an exposure in a specific study. In this section, we describe two methods for adjusting for confounders under the assumption of no unmeasured confounders. Others are discussed in chapter 5. In addition to these methods, we describe a method that is well established to control for unmeasured confounders in both RCTs and observational studies. This methodology is called the instrumental variables (IVs) method and we will describe it in a separate section.

## 3.1 Stratification

In order to estimate the causal effect of exposure $X$ on outcome, one has to account for the effect of confounding factors. Confounding adjustment may improve precision,

or make precision worse, but is primarily designed to avoid bias. One method to deal with confounding is by stratification. If the no unmeasured confounders assumption

$$Y_x \perp\!\!\!\perp X|Z \tag{2.3}$$

holds, then

$$f(y_x|Z) = f(y_x|X = x, Z) = f(y|X = x, Z),$$

and

$$\mathrm{E}(Y_x|Z) = \mathrm{E}(Y_x|X = x, Z) = \mathrm{E}(Y|X = x, Z).$$

That is, the density and mean of $Y_x$ can be obtained by dividing the population within strata defined by the values of $Z$ and estimating the outcome density and mean in that stratum for subjects with $X = x$. This approach is called stratification and it applies often when confounding factors are categorical. Stratification involves an unknown data law, $f(y|X = x, Z)$ which can be estimated non-parametrically. The resulting causal effect estimator may behave erratically in moderate sample size because $Z$ is usually high dimensional. Suppose that we will postulate the following conditional model

$$\mathrm{E}(Y_x|Z) = g\{\omega(Z) + \psi^* x\} \tag{2.4}$$

where $g(.)$ is a known link function and $\omega(.)$ known up to a finite dimensional nuisance parameter for each $x$. This model is called a conditional structural mean model. For instance, when $g$ is the identity function and $\omega(Z)$ linear, model (2.4) is the following causal model,

$$\mathrm{E}(Y_x|Z) = \alpha_0^* + \alpha_1^* Z + \psi^* x$$

where $\alpha_0^*$, $\alpha_1^*$, and $\psi^*$ are unknown finite dimensional parameters. The causal parameter equals $\psi^* = \mathrm{E}(Y_1 - Y_0|Z)$. If $\psi^*$ is non-constant in $Z$, then $Z$ is said to be a moderator of the effect of $X$ on $Y$. If there is a linear interaction between exposure and confounders, the structural model for the conditional mean of $Y_x$ given $Z$ is as follows

$$\mathrm{E}[Y_x|Z] = \alpha_0^* + \alpha_1^* Z + \psi_1^* x + \psi_2^* xZ$$

where $\psi_2^*$ is an interaction effect of exposure and confounders. The causal parameter is then equal to $\psi_1^* + \psi_2^* Z = \mathrm{E}(Y_1 - Y_0|Z)$. An example of interaction is seen in the case of oral contraceptive use, $X$, and its effect on cardiovascular disease, $Y$, adjusted for smoking, $Z$. Because smoking $Z$ amplifies thromboembolic-disease risk $Y$ in oral contraceptive users, interaction is said to exist (Curtis et al., 2006). This is why oral contraceptives carry a boxed warning advising against their use in smokers. In this

model, $Z$ is a moderator of the effect of $X$ on $Y$ if $\psi_2^*$ is non zero. In the case where the outcome $Y$ is dichotomous, model (2.4) is for instance as follows,

$$\text{logit}P(Y_x = 1|Z) = \alpha_0^* + \alpha_1^* Z + \psi^* x$$

where $\text{logit}(p) = \log\{p/(1-p)\}$. The causal odds ratio is then

$$\exp(\psi^*) = \frac{P(Y_1 = 1|Z)/P(Y_1 = 0|Z)}{P(Y_0 = 1|Z)/P(Y_0 = 0|Z)}.$$

Under assumption (2.3), the conditional SMM (2.4) implies the following model for the observed data

$$\text{E}(Y|X, Z) = g\{\omega(Z) + \psi^* X\}, \tag{2.5}$$

for instance,

$$\text{E}(Y|X, Z) = \alpha_0^* + \alpha_1^* Z + \psi^* X.$$

Thus consistent estimates for $\psi^*$ under the conditional SMM (2.5) can be obtained by fitting an ordinary conditional mean model, or for instance, by fitting a logistic conditional mean model

$$\text{logit}P(Y = 1|X, Z) = \alpha_0^* + \alpha_1^* Z + \psi^* X.$$

## 3.2 Propensity score

In model (2.4) the association $\omega(Z)$ of confounders with outcome is often problematic to specify. Because estimates for $\psi^*$ may then be inconsistent, tests of no causal effect may be invalid. When the assumed model for $\omega(Z)$ is incorrect and exposure $X$ is dichotomous, one solution is to adjust for the propensity score (Robin, 1997; Rosenbaum and Robin, 1983; Rosenbaum, 2002). This is an approach which indeed provides a means for adjusting for selection bias in observational studies of causal effects. Creating strata in which subjects are matched on the propensity score allows one to balance measured variables between treated and untreated subjects. The propensity score is the conditional probability of receiving a given exposure (treatment) given a vector of measured covariates. It is usually estimated using logistic regression. Therefore instead of the high dimensional confounder $Z$, one adjusts for the propensity score

$$p(Z) = P(X = 1|Z)$$

where p(Z) is the propensity score and $p$ encodes for the propensity score. The key point is that the no unmeasured confounders assumption implies that

$$Y_x \perp\!\!\!\perp X|p(Z).$$

Because

$$\begin{aligned} P(X=1|Y_x, p(Z)) &= \mathrm{E}\{P(X=1|Y_x, Z)|Y_x, p(Z)\} \\ &= \mathrm{E}\{P(X=1|Z)|Y_x, p(Z)\} \\ &= p(Z). \end{aligned}$$

Therefore, it follows that adjusting for the propensity score is sufficient. For example when

$$\mathrm{E}(Y_x|p(Z)) = \alpha_0^* + \alpha_1^* p(Z) + \psi^* x$$

consistent estimates of $\psi^*$ can be obtained by fitting the following model for the observed data

$$\mathrm{E}(Y|p(Z)) = \alpha_0^* + \alpha_1^* p(Z) + \psi^* X.$$

The propensity score adjustments work better in settings where there is little overlap in the distribution of confounders between exposed and unexposed subjects (Rubin, 1997; Lunceford and Davidian, 2004). When exposure $X$ is continuous, estimates for the parameter $\psi^*$ indexing model (2.4) may be obtained via G-estimation (Robins et al., 1992; Brumback et al., 2003) which is discussed in chapter 5.

**Example 1**. Birth weight is of concern to physicians because it is a strong predictor of infant mortality and birth defects. Women‘s behavior during pregnancy (including diet, smoking habits and receiving prenatal care) can greatly alter the chances of delivering a baby of normal birth weight. We reanalyze birth weight data (Hosmer and Lemshow, 2000) of 189 babies. Table 2.2 shows the estimates of causal effect of

Table 2.2: *Estimates and 95% confidence intervals and p-value of the smoking effect with different methods for adjusting for confounders. The subtable left is presented when the outcome is birth weight measured in gram units and the subtable right is when the outcome is an indicator of low birth weight.*

| | Continuous outcome | | | Dichotomous outcome | | |
|---|---|---|---|---|---|---|
| Adjusting | $\hat{\psi}$ | 95 % CI | P.value | $\exp(\hat{\psi})$ | 95 % CI | P.value |
| No adjusting | -0.28 | (-0.49, -0.07) | 0.0090 | 2.02 | (1.08,3.78) | 0.0270 |
| OLS | -0.35 | (-0.59, - 0.13) | 0.0007 | 2.88 | (1.30, 6.22) | 0.0070 |
| PS | -0.36 | (-0.57 , -0.16) | 0.0022 | 2.94 | (1.35, 6.38) | 0.0060 |

maternal smoking during pregnancy on average birth weight following different methods for controlling confounding. In this example, the confounders are: maternal age, maternal weight at last menstrual period, race, and maternal history of hypertension. The percentage of self reported nonsmokers was 61 %. The adjustment for confounding was done twice, once for the case where outcome (birth weight) is continuous (subtable left) and once where outcome is dichotomous (low birth weight) (subtable right). Table 2.2 shows, average causal effects of smoking on average birth weight are relatively similar for the OLS and PS causal effect estimators (see subtable left) and suggests strong evidence of a harmful smoking effect. Specifically, the PS estimator suggests a 360 gram (95% confidence interval $(160, 570)$) decrease in average birth weight due to smoking.

# 4 Randomized clinical trials and the problem of noncompliance

In a typical clinical trial, patients are mainly randomized to one specific dosing strategy of treatment $A$ or treatment $B$ over a period of time. The primary analysis should compare patients in their randomly assigned treatment groups $A$ and $B$. There is almost universal agreement among clinical trialists that such intention-to-treat (ITT) analysis is the primary analysis of randomized clinical trials (RCTs).

**Intention-to-treat (ITT) analysis**. ITT analysis is a standard analysis in randomized clinical trials (RCTs). This approach analyzes subjects by comparing their outcomes according to the intention to treat each subject rather than treatment received. In fact, ITT estimators measure the effect of assignment rather than treatment. For instance, in a RCT, the randomization indicator $R_i$ is equal to 1 if subject $i$ is assigned to treatment and 0 if he/she is assigned to placebo; then $\hat{\psi}_{ITT} = \hat{\mathrm{E}}(Y_i|R_i = 1) - \hat{\mathrm{E}}(Y_i|R_i = 0)$. This is the most suitable approach for pragmatic trials that aim to measure the effectiveness (the benefit of treatment policy) of a treatment and is less suitable to measure the efficacy (the benefit of actually receiving a treatment) of a treatment (e.g., Sommer and Zeger, 1991; Goetghebeur et al., 1998; Dunn et al., 2005).

For scientific experiments involving human participants, noncompliance and partial compliance are very common. Noncompliance might generally be defined as partial or zero adherence to a subject's prescribed treatment. The term 'noncompliance' or 'non-adherence' is commonly applied for these events and the term 'departure from randomized treatments' is also often used. In practice, it is reasonable to believe that most, if not all, randomized clinical trials are not an ideal experiment where sub-

jects actually receive their randomly allocated treatment exactly as they have been prescribed in the experiment's treatment protocol. This is particularly true for complex treatments (Dunn and Bentall, 2007). Because many patients may fail to take their prescribed medicine, they would not actually receive all of the prescribed dose they were randomly assigned to (Efron and Feldman, 1991; Goetghebeur et al., 1998; Nagelkerke et al., 2000; Goetghebeur and Vansteelandt, 2005; Dunn et al., 2005). Others may take the medicine or receive an alternative treatment that had been allocated to other people in the trial. The reasons for noncompliance may be associated with the outcome variable and thereby act as measured confounders (prognostic variables), more importantly, one typically also needs to acknowledge the existence of unmeasured confounders (Efron and Feldman, 1991; Fischer-Lapp and Goetghebeur, 1999; Nagelkerke et al., 2000). Noncompliance with assigned treatment therefore is a ubiquitous issue in RCTs. There are no universally accepted guidelines on noncompliance analysis to date. The following examples show different ITT estimates in imperfect RCTs.

**Example 2**. Sommer and Zeger (1991) study the effect of vitamin A supplementation on mortality in preschool children in a trial carried out in Indonesia. Children allocated to the vitamin A supplementation did not always receive it and no placebo treatment was given in the control group due to a local law. In totally 23 682 children were randomized to receive vitamin $A$ supplementation or no intervention (in a cluster randomized fashion) (Table 2.3). Ignoring the clustering, ITT analysis suggests

Table 2.3: *Vitamin A trial data, according to Sommer and Zeger (1991). Dichotomous outcome (alive/dead).*

| | Treatment compliance | | | Control compliance | | |
|---|---|---|---|---|---|---|
| | No | Yes | Total | No | Yes | Total |
| Alive | 2385 | 9663 | 12048 | - | - | 11514 |
| Dead | 34 | 12 | 46 | - | - | 74 |
| Total | 2419 | 9675 | 12094 | - | - | 11588 |

an effectiveness with relative risk, $RR = 0.60$ and 95 % CI $(0.41, 0.86)$. That is, allocation to vitamin $A$ reduced mortality to 60%.

**Example 3**. Goetghebeur et al. (1998) present further details on this trial of the effect of vitamin A supplementation on mortality in preschool children on a slightly modified data base. They consider that children were randomized to receive two doses of vitamin A, one at the beginning of period 1 (0-4 months) and one at the beginning

of period 2 (5-8 months) or not, any child surviving 12 months was considered to be censored. Compliance can be to either the first dose, the second dose, or both or neither. Specifically, they refine the binary compliance summary, retaining the number of high dose vitamin A pills children received: 0, 1, or 2. By using Cox regression, ITT analysis suggests a moderate but potentially important effect ($\beta = -0.43$) and 95 % CI $(-0.73, -0.23)$ of vitamin $A$ supplementation on mortality. (Note that in example 1, the causal parameter is a risk ratio, and in this example it is the effect of measured as a hazard ratio.)

**Example 4**. Cole and Chu (2005) present the Herpetic Eye Disease Study which randomized 703 ocular herpes patients to 365 days of acyclovir or placebo between 1992 and 1996. In this trial, there was over 90 per cent compliance in both arms. Compliance was assessed by pill counts at study visits planned for 1, 3, 6, 9 and 12 months post-randomization. If a patient did not return pill bottles, then compliance was estimated from medication cards used by the patient to record when pills were taken and/or the physician's judgment. The result was an estimate of the percent compliance over the intended follow-up, irrespective of endpoint status (i.e., count of treated days divided by 365). The hazard of recurrence in the acyclovir arm was 0.55 times the hazard in the placebo arm using ITT analysis with 95 % CI $(0.41, 0.75)$ (Data not presented here).

The limitations of an analysis which ignores noncompliance have received growing concern in the causal inference literature. It is important to attempt to reduce the rate of noncompliance in these circumstances but at the same time one wishes to estimate the causal effect of exposure at the levels that were actually taken. Indeed, the challenge to the trialists is to draw valid inference from the data actually obtained from such a trial. An alternative to the ITT analysis is the 'as-treated analysis', in which subjects are classified by the treatment actually received. As argued before, this analysis is unsatisfactory since confounders associated with switching treatment threaten the causal interpretation of treatment effects (Nagelkerke et al., 2000). For example, if the most treatment resistant patients in the less effective treatment group switch to the more effective treatment, they may decrease the average level of improvement for the more effective treatment, also leading to an underestimate of the true causal difference between the two treatments. An other analysis that most authors also use is the 'per-protocol analysis' (pp), which excludes any data collected from a subject after they have departed from the randomized treatment. In words, subjects who depart from randomized treatment are included in the analysis only up until the point of departure (White, 2005).

All of these approaches consider treatment actually received in combination with treatment assigned, two jointly observed variables. A very different approach moves beyond this. In the placebo controlled noncompliance context (Angrist et al., 1996;

Dunn et al., 2005), subjects are often classified into four categories: (a) compliers, are those who take the treatment whatever treatment is assigned (b) always-takers, are those who would take the assigned active treatment no matter what their assignment (c) never-takers, are those who would not take the treatment no matter what their assignment, and (d) defiers, are those who would take the treatment opposite from their assignment regardless of the treatment they were assigned to. In the placebo controlled clinical trials one often assumes that the subjects allocated to the placebo do not get access to the treatment (no contamination) and there is only one form of noncompliance to be observed. When there is no noncompliance in the control group (no contamination), subjects are divided into compliers and never-takers. Treatment effect is most meaningful for compliers. With no noncompliance in the control group, efficacy among compliers equals efficacy among the treated. In the next example, we compare the ITT, as-treated and PP-analysis on vitamin $A$ trial.

**Example 5**. ITT analysis showed a moderate effectiveness with relative risk ($RR = 0.60$) and 95 % CI $(0.41, 0.86)$. That is, allocation to vitamin $A$ reduced mortality to 60% in example 2. By performing a per-protocol analysis, we find a RR of 0.19 with 95 % CI $(0.11, 0.36)$, which shows mortality was much higher among children randomized to vitamin $A$ who did not receive it than in the control arm. By performing an as-treated analysis, we find a RR of 0.16 with 95 % CI $(0.09, 0.29)$, which shows an even more extreme risk ratio.

Therefore, if investigators were to take a naive look at the association between measured treatment compliance and outcome in the treated group they would most likely be misled. In view of this, there have been a number of recent proposals on how to use compliance and how to adjust for noncompliance in the analysis. Briefly, some authors used treatment arm (randomization) as an instrumental variable for correcting noncompliance in trials comparing either several levels of a single active treatment or a single active treatment to placebo (e.g., Robins, 1989; Angrist et al., 1996; Goetghebeur et al., 1998). Others (e.g., Rubin, 1997) proposed maximum likelihood and Bayesian inferential methods for CACE, which impose additional assumptions and are then more efficient than standard instrumental variable methods. To reduce the noncompliance impact, some authors also resist analyzing the data in a double-blind manner when subjects in the placebo arm lack access to treatment. However, these approaches generally require various strict assumptions. Efron and Feldman (1991) show compliance is an uncontrolled covariate. They "modeled the mean effect of treatment (cholestyramine) in a placebo controlled trial as a linear function of the percentage of assigned cholestyramine that is actually taken". Recent methods for causal inference offer new insights and opportunities by estimating what ITT analysis would have been, had everyone complied perfectly. Robins (1994), and Goetghebeur and Lapp (1997) have used semi-parametric structural linear mean models (defined in

next section) for compliance adjustment in RCTs. These models consider treatment-free outcome as a reference for observed outcome.

# 5 Instrumental variables (IVs)

The method of instrumental variables (IV) has originally been applied in the econometrics literature when the independent variable $X$ is correlated with the error term $\epsilon$ in the linear regression of $Y$ on $X$. Suppose that the outcome $Y$ is linearly related to the observed exposure $X$ as follows

$$Y = \beta_0^* + \beta_1^* X + \epsilon. \tag{2.6}$$

A consistent OLS estimator for $\beta^*$ can be obtained when $\epsilon$ and $X$ are uncorrelated (Martens et al., 2006; Hernán and Robins, 2006). Now suppose that in the data setting there is an unmeasured confounder $U$, such that the true model in fact is

$$Y = \beta_0^* + \beta_1^* X + \gamma^* U + \delta \tag{2.7}$$

where $\delta$ is the random deviation of $Y$ from the expected value based on the model and is independent of $U$ and $x$, but we still proceed with our naive OLS estimator as for model (2.6). Since $X$ and $\epsilon$ are no longer uncorrelated, we will obtain an inconsistent estimate for $\beta^*$. In view of this problem, a much more common approach in econometrics traditionally and recently in epidemiological applications, is to find a variable that is strongly correlated with exposure $X$, has no direct effect on the outcome $Y$, and is conditional independent of $Y$ given $X$. Such a variable is called an instrumental variable (IV). Suppose that the random variable $V$ is an instrumental variable, then

$$X = \theta_0^* + \theta_1^* V + \tau$$

where error terms $\epsilon$ and $\tau$ are uncorrelated and $\theta_1^* \neq 0$. An asymptotically unbiased estimate of the effect of $X$ on $Y$, $\beta_1^*$, can then be obtained as

$$\hat{\beta}_1^* = \frac{\sum_{i=1}^{n}(V_i - \bar{V})(Y_i - \bar{Y})/n}{\sum_{i=1}^{n}(V_i - \bar{V})(X_i - \bar{X})/n}.$$

This estimator is in fact consistent, when $V$ is uncorrelated with the error term $\epsilon$ (Hernán and Robins, 2006). We can say that the numerator measures the effect of the IV on the outcome $Y$ and the denominator measures the effect of the IV

on the exposure $X$. This estimator is indeed obtained by a two-stage least squares algorithm. First regressing $X$ on $V$, and then regressing $Y$ on the predicted value of $X$ obtained from the first regression of $X$ on $V$ (Wooldridge 2002, 2003). In the case of a dichotomous IV, the numerator and denominator are simply the difference in mean outcome and mean exposure between $V = 1$ and $V = 0$, respectively.
Random variable $V$ is generally an instrumental variable (IV) if:

- it is associated with exposure $X$;
- it affects the outcome $Y$ only through $X$;
- the association between $V$ and $Y$ is unconfounded.

Instrumental variable methods were invented more than 70 years ago, and were first used in economics and econometrics in connection with structural equation models. They have more recently entered the medical, epidemiological and biostatistical literature, mainly in connection with noncompliance adjustment in randomized clinical trials (Goetghebeur et al., 1998; Robins, 1998; Greenland, 2000) and in the context of Mendelian randomization studies (Didelez and Sheehan, 2007; Lawlor et al., 2008; Didelez et al., 2008), as well as in observational studies. Although, randomized controlled trials (RCTs) remain the gold standard for research, the analysis based on these trials is problematic when in the presence of noncompliance there are unmeasured confounders (Robins, 1998; Goetghebeur et al., 1996; Greenland, 2000; Dunn, 2005; Hernán and Robins, 2006). However, the crux of observational studies is also that treatment status is not controlled by the researcher and can be related to various background variables (confounders) which might be unknown (Rosenbaum and Rubin, 1983; Rubin, 1997). Observational studies compare subjects that have chosen (or whose physicians have chosen) to take a treatment of interest to the other group of subjects. Any difference in outcome between the two groups could be due to the treatment altering the outcome or it could be that the treatment has no effect but the reason for choosing to take the treatment is the cause of any difference. In view of this, the instrumental variable (IV) method is one methodology that has been used to overcome the effect of unmeasured confounders in both types of studies. An instrumental variable (IV) must satisfy some strong assumptions to allow for consistent effect estimates. Indeed, its power is derived solely from the assumption that the instrument only affects the outcome indirectly through the exposure (Martens et al., 2006; Hernán and Robins, 2006). Although this assumption is untestable based on the data only, it is plausible by design. The main problem therefore remains to find a convincing instrumental variable that has strong correlation with exposure $X$. The stronger this association the more precise the instrumental variable estimate of exposure effect.

## 5.1 IVs and imperfect randomized studies

In RCTs, treatment assignment can provide a perfect IV for the effect of received treatment on outcome and indeed a randomization based exposure adjusted analysis is an alternative to ITT analysis (Greenland, 2000; Goetghebeur and Vansteelandt, 2005; Dunn, 2005; Hernán and Robins, 2006). The dichotomous randomization variable $R$ takes the value 1 if a subject is randomly allocated to receive the active treatment or new intervention, and the value 0 if the allocation is to receive the placebo treatment or control condition. For ease of discussion, we consider the dichotomous treatment $X$ in which $X = 1$ for exposed, 0 otherwise. The outcome variable can be continuous (often normally distributed) or dichotomous. The variable $Z$ is a covariate measured prior to randomization (pre-randomization). The potential randomized assignment outcome is defined as $X_r$ for $r = 0, 1$. Then the effect of randomization on receipt of treatment for subject $i$ is: $X_{i1} - X_{i0}$. We cannot observe this difference for each subject because that subject is only ever randomized to one of the conditions. The treatment status can then be expressed as $X = (1 - R)X_1 + (1 - R)X_0$. In practice, we observe $X$. Following Angrist et al. (1996), there are only three possibilities for the randomization effect on $X$: $X_1 - X_0 = 1$, or 0, and or $-1$. These are corresponding to the compliers; always-takers, never-takers; and defiers, respectively where, $X_1 - X_0 = -1$ (defiers) is often not realistic and ignored as a possibility in the analysis. In this setting, $Y_{irx}$ may represent the counterfactual outcome of interest that would have been observed if (possibly contrary to fact) subject $i$ were randomized to arm $r$ and treatment $x$ were given; $r = 0, 1$, $x = 0, 1$ (Robins, 1998; Vansteelandt and Goetghebeur, 2003). For instance in the vitamin $A$ trial (Sommer and Zeger, 1991), $Y_{i11}$ would have been observed if child $i$ were randomized to vitamin $A$ and vitamins $A$ were received. Consider now a double-blind RCT with noncompliance where the relationship between received treatment and outcome is influenced by unobserved confounders. To be able to infer the causal effect of $X$ on $Y$, the following assumptions must be satisfied,

1. $R$ is associated with exposure $X$;

2. $R$ affects the outcome $Y$ only through $X$ (exclusion restriction);

3. no confounding for the effect of $R$ on $Y$ (randomization assumption).

With the above assumptions randomization $R$ is an instrumental variable (IV) for the effect of received treatment on outcome. Let $U$ indicate confounding factors, which may be partly observed or entirely unobserved, that affect both the exposure and the outcome. Figure 5 presents a DAG that satisfies these assumptions. These

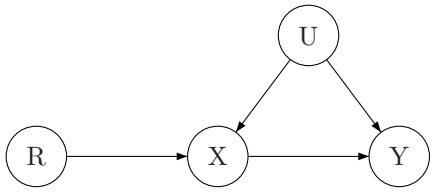



Figure 5: *A double-blind randomized trial with randomization assignment $R$ as an IV. $U$ is partly observed or entirely unobserved confounders.*

assumptions cannot easily be tested from the observable data as they involve the unobservable $U$. However, in a double-blind RCT the above assumptions are satisfied by design (Hernán and Robins, 2006). Assumption (1) holds because trial subjects are more likely to receive treatment if they were assigned to treatment. Assumption (2) holds by effective double-blindness. This assumption indeed says that for essentially all subjects, assigned treatment has no direct effect on the outcome when received treatment is held fixed. Thus, under this assumption $Y_{rx} = Y_x$. The assumption (3) holds by the random assignment. The assumption (3) implies that $R \perp\!\!\!\perp Y|X, U$. In addition to the above assumptions, the monotonicity assumption $P(X_1 > X_0) = 1$ may also be assumed to hold. This rules out the existence of defiers and classifies a subject of the population into always-takers, compliers, and never-takers. The IV assumptions are also made explicitly conditional on covariates to allow for the fact that instruments can be related to these background variables. The usual instrumental variable estimator in a randomized trial with noncompliance may be viewed as estimating the differential effect of treatment on compliers. It can be simply written as

$$\hat{\psi}_{IV} = \frac{\hat{\mathrm{E}}(Y|R=1) - \hat{\mathrm{E}}(Y|R=0)}{\hat{\mathrm{E}}(X|R=1) - \hat{\mathrm{E}}(X|R=0)}.$$

As stated, the numerator is the standard estimator for the treatment effect of $R$ on $Y$ (ITT of $R$ on $Y$), and the denominator is the standard estimator for the treatment effect of $R$ on experimental exposure $X$ (ITT of $R$ on $X$). Because $R$ is randomly assigned, as in the case of a natural experiment, both estimates are consistent (Greenland, 2000). It follows from $\hat{\psi}_{IV}$ that, the weaker the association between $R$ and $X$ the more the ITT effect will be inflated because of the shrinking denominator.

Consider now treatment effects which are linear in received dose

$$Y_x - Y_0 = \psi^* x$$

then

$$Y = Y(X) = Y_0 + \psi^* X.$$

Due to randomization $R$ is an IV, it is not associated with outcome in the absence of treatment, that is, not associated with the potential treatment-free outcome. Because $Y_0$ is independent of $R$ given $Z$, by the randomization assumption of the IV

$$\begin{aligned} \mathrm{E}(Y|R,Z) &= \mathrm{E}(Y_0|R,Z) + \psi^* \mathrm{E}(X|R,Z) \\ &= \mathrm{E}(Y_0|Z) + \psi^* \mathrm{E}(X|R,Z). \end{aligned}$$

Assuming for instance models: $\mathrm{E}(X|R,Z) = \alpha_0^* + \alpha_1^* R + \alpha_2^* Z$, fit this model first to obtain predictions $\mathrm{E}(X|R,Z;\hat{\alpha})$. At the second stage another linear model $\mathrm{E}(Y_0|Z) = \beta_0^* + \beta_1^* Z$ allows to estimate $\beta^*$. Then an estimate of $\psi^*$ is obtained by fitting the model

$$\mathrm{E}(Y|R,Z) = \beta_0^* + \beta_1^* Z + \psi^* \mathrm{E}(X|R,Z;\hat{\alpha})$$

for instance using ordinary least squares. Thus, the IV estimators for $\psi^*$ is derived in two-stages:

1. regress compliance on the instrumental variable (and covariates) and obtain predictors;

2. regress outcome on predicted compliance (and covariates).

Standard errors must acknowledge the fact that predictions are imprecise. Note that, standard software for Two-stage least squares estimation thus allows to correct for noncompliance under certain assumptions.
We present a Blood Pressure Reduction trial as an example of a randomized trial and analyze it in this chapter and in chapter 4.

## Blood Pressure Reduction trial

Goetghebeur and Lapp (1997), and Goetghebeur and Vansteelandt (2005) analyzed a double-blind, three-arm trial on blood pressure reduction which randomized some 300 hypertensive patients in the U.K. in the years 1989-1990. After a initial run-in period of 4 weeks on placebo, patients were randomly assigned to one of two experimental treatments (Nebivolol, Atenolol), or placebo. Each treatment is prescribed at one tablet a day at breakfast. At study entry, baseline characteristics such

as sex, height, weight, age and diastolic blood pressure (DBP), were recorded. In all, five clinic visits were scheduled a fortnight apart, starting from the beginning of the run-in. DBP was measured at each visit. At visit 3, the 4-week active treatment period started. For a sample of 164 patients, this indicated the start of electronic compliance monitoring: a device in the pill container measured exact times at which patients opened and closed the container. Because of technical problems with the electronic monitors, compliance data of 10 patients were missing. This left samples of size 54, 49, and 51 available for analysis on treatment arms Nebivolol, Atenolol, and the placebo arm respectively. The compliance measure used is the percentage of assigned active dose which was actually taken.
Table 2.4 compares the results of an ITT analysis, As-treated analysis, and IV estimators (with and without adjustment for covariates) in the DBP trial. For instance, IV estimator I in the left subtable shows that among subjects who chose to take on average 1 pill (Nebivolol) per day, the blood pressure reduction would on average have been 9.56 mmHg (95 % CI (5.29, 13.83)) smaller had they not taken any pills (Nebivolol). The subtable right shows also how with different observed exposures a significantly different average causal effect on the blood pressure is found.

Table 2.4: *Estimates and 95% confidence intervals and p-value of the effect of both treatments (Nebivolol and Atenolol) versus placebo. The IV estimator I indicates that baseline covariates are not adjusted and the IV estimator II indicates that all baseline covariates are adjusted for DBP trial.*

| | Nebivolol versus Placebo | | | Atenolol versus Placebo | | |
|---|---|---|---|---|---|---|
| Estimators | $\hat{\psi}^*$ | 95 % CI | P.value | $\hat{\psi}^*$ | 95 % CI | P.value |
| ITT | -8.11 | (-11.66, -4.56) | $10^{-4}$ | -9.55 | (-13.77, -5.34) | $2\times10^{-4}$ |
| As treated | -5.98 | (-11.37, -3.96) | $2\times10^{-4}$ | -8.54 | (-12.82, -4.25) | $10^{-4}$ |
| IV estimatorsI | -9.56 | (-13.83, -5.29 ) | $2\times10^{-4}$ | -10.38 | (-15.06, -5.69) | $3\times10^{-4}$ |
| IV EstimatorsII | -7.63 | (-11.77, -3.48 ) | $2\times10^{-4}$ | -7.44 | (-11.92, -2.96) | $10^{-4}$ |

For instance, IV estimator I suggests that the average blood pressure reduction would have been 9.56 mmHg (95% CI $5.29, 13.83$) smaller over the study period among those who choose to take on average one pill per day, had they not taken the exposure. In contrast, ITT shows 8.11 mmHg (95% CI $4.56, 11.66$). Greenland (2000) obtains the IV estimation for the causal effect of vitamin $A$ on mortality for the data in Table 2.5, where he applies the IV method to control for confounding when acknowledging noncompliance in this trial. First, the proportion of subjects who are compliers is estimated in the treatment arm: $\hat{P}_c = 0.80$ (c indicates compliers). He then obtains the IV estimate of the risk difference (RD) produced by treatment comparing $x = 1$

to $x = 0$: $\frac{\hat{m}_{\bullet 1} - \hat{m}_{\bullet 0}}{\hat{p}_c} = -324$ per 100 000=-0.324 % with 95 % CI (-0.55 %, -0.10 %), this is a 324/639=51 % risk reduction. The quantities $m_{\bullet 1}$ is the average outcome if every one is assigned treatment (R=1): $\hat{m}_{\bullet 1} = \frac{46}{12095} = 380$ per 100 000; and $m_{\bullet 0}$ is the average outcome if every one is assigned to placebo (R=0): $\hat{m}_{\bullet 0} = \frac{74}{11588} = 639$ per 100 000. Because the original data set is not available, this estimated CI is not exact for the Sommer and Zeger data set (Greenland, 2000).

Table 2.5: *One year mortality data from the randomized trial of vitamin A supplementation in children (Sommer and Zeger, 1991). Risk indicates deaths per 100 000 within one year.*

| | $R = 1$ | | | $R = 0$ | |
|---|---|---|---|---|---|
| | $X = 1$ | $X = 0$ | Total | $X = 1$ | $X = 0$ |
| Deaths ($Y = 1$) | 12 | 34 | 46 | 0 | 74 |
| Total | 9675 | 2419 | 12094 | 0 | 11588 |
| Risk | 124 | 1406 | 380 | undefined | 639 |

## 5.2 IVs in observational studies

The interest in IV methods stems from the fact that, in the presence of unmeasured confounders, one can estimate the causal effect of nonrandomized exposure on an outcome of interest by using an instrumental variable. Specifically, while the causal effect of exposures are distorted by unmeasured confounders, which is a major concern of researchers carrying out epidemiological and Mendelian randomization studies, the instrumental variable (IV) approach is relatively straightforward once an IV has been identified and is strong enough to correct for the bias confounding in observational studies. Finding a valid IV or an IV which is strongly correlated with the exposure is however quite difficult. We list some convincing examples. In a study of the risk of gastrointestinal (GI) complications attributable to different non-steroidal anti-inflammatory drugs (instead of Cox-2 inhibitors), Brookhart et al. (2007) choose the physician's prescribing preference as an instrumental variable and argued it might satisfy the IV assumptions (see chapter 3). Leigh and Schembri (2004) use the cigarette price per region as an IV to estimate the effect of smoking on health. The use of genetic variants as instrumental variable (IV) is applied to analyze the causal relationship between exposure and a disease outcome. Some papers have described Mendelian randomization as an IV analysis (Thomas and Conti, 2004; Didelez

and Sheehan, 2007; Lawlor et al., 2007). In the next chapter, we will elaborate on the application of IV methods for handling this long-standing inferential problem in observational studies. Before doing so we point to one final important complication in this filed: measurement error in exposure.

# 6 Measurement error and causal models

In addition to confounding, an important but largely overlooked impediment using compliance measurements to validly estimate the effect of received assigned drug dosing, is measurement error. Although in recent years, new measuring devices such as drug concentrations and electronic caps that monitor the opening and closing of a medication container have been used besides pill counts and questionnaires to measure compliance, measurement error remains a serious problem. Medication event monitoring systems (MEMS) record the exact timing of opening of drug containers, providing more informative data than previously available (Urquhart and De Klerk, 1998). Nonetheless, accurate measures of treatment compliance have proved difficult to obtain (Dunn, 1999; Goetghebeur and Vansteelandt, 2005). Partly because each method only indirectly measures drug intake and none of them records whether a subject swallows tablets or not. Moreover, there are concerns that subjects may take more than 1 pill out of the MEMS container, or they may not swallow all pills that they took out. In practice, it is then reasonable to believe that the treatment compliance measurements are measured with error, that is, true treatment compliance measurements may inherently differ from the observed treatment compliance measurements. More specifically, the true exposure (treatment compliance measurement) will be equal to the observed exposure plus or minus some error value. It is well known that (as discussed in chapter 1) not adjusting for systematic measurement error can downwardly or upwardly bias estimates for the causal effect of exposure. Therefore, systematic measurement error in exposure is a real concern in many practical settings and the causal effect estimator in the context of compliance measurements continues to be asymptotically biased when systematic measurement error in exposure is ignored. While a large literature has become available on measurement error in explanatory variables for the linear and nonlinear association models (as discussed in the chapter 1), little attention has been paid so far to insight into the impact of the measurement error problem in exposure in the context of newly developed causal models with compliance measurements. “Dunn (1999) is one of the rare authors who recognizes explicitly the importance of exposure measurement error in the field of causal inference for compliance data. “He points out that all clinical measures are fallible, and that compliance measures are particularly hard to obtain”. Goetghebeur

and Vansteelandt (2005), have investigated this problem and show how this can be handled when the average size of the error is known. In next the sections, we review, linear structural mean models (LSMMs) (Robins, 1994; Goetghebeur and Lapp, 1997; Goetghebeur and Vansteelandt, 2005) for the analysis of randomized trials subject to noncompliance and consider for their performance under common complications of measurement error in exposure. We provide a worked example of applying LSMMs with a correction for measurement error in exposure in the Blood Pressure Reduction trial (Goetghebeur and Vansteelandt, 2005). In this context, we also consider log-linear and logistic structural mean models (Robins, 1997; Vansteelandt and Goetghebeur, 2003; Robins and Rotnitzky, 2004; Goetghebeur and Vansteelandt, 2005).

## 6.1 Structural mean models

The linear, log-linear and logistic structural mean models (SMMs) (Robins, 1994, 1998; Fischer-Lapp and Goetghebeur, 1999; Vansteelandt and Goetghebeur, 2003; Goetghebeur and Vansteelandt, 2005) are recently developed causal models that allow to adjust for noncompliance to the prescribed therapy in randomized clinical trials when one is interested in the effect of the amount of drug actually taken. The structural mean models (SMMs) express the expected causal effect of treatment as a function of the amount of drug actually taken. In the next section, we introduce LSMMs and apply then to reanalyze the Blood Pressure Reduction trial to estimate the effect of the percentage of assigned active dose that subjects actually took. We investigate the impact of measurement error on estimated exposure effects. Specifically, we examine the impact of measurement error under a plausible range of expected errors on the measurement of both active treatments (namely, Nebivolol and Atenolol) in this trial by a sensitivity analysis as in by Goetghebeur and Vansteelandt (2005).

## 6.2 Linear structural mean models (LSMMs)

Consider the Blood Pressure Reduction trial described in the previous section. Having defined potential or counterfactual outcomes and causal effects in section 2.1, we now formulate these under the LSMM. We develop this for a single active treatment versus placebo. We let the received experimental treatment be continuous, for this example, and denote the percentage of active drug actually taken for subject $i$ by $X_i$. Let $C_i$ denote the observed compliance pattern under the assigned treatment. Then, on the active treatment arms, $X_i = h(C_i)$ is a meaningful summary of the pattern of exposure (Goetghebeur and Lapp, 1997). In this trial there is no contamination expected. The randomization indicator $R_i$ is to equal 1 if randomized to the treatment arm and 0 to placebo arm. We denote $Y_i$ as the observed outcome

variable. To estimate the causal effect of $X_i$ on the outcome, we would ideally like to measure how subjects on the active treatment arm would respond if they took placebo instead of active treatment. As stated, the reference outcome for this possible scenario is called a potential treatment-free outcome and denoted by $Y_{i0r}$, assignment treatment following $R_i = r(r = 0, 1)$. With this notation we are making the exclusion restriction assumption (Angrist et al., 1996) that randomization assignment $R_i$ has no direct effect on outcome. In a double-blind trial, one expects any effect of randomized assignment on outcome to be captured by the effect of assignment on exposure. Under this assumption, we can use $Y_{i0}$, $(Y_{i0,0} = Y_{i0,1} = Y_{i0})$, to denote the potential treatment-free outcome of patient $i$ regardless of treatment arm. Thus, the contrast $Y_i - Y_{i0}$ is the causal effect of actually received of exposure $X_i$ for subject $i$ relative to no exposure. LSMMs regress the mean of this contrast at fixed levels of exposure $X_i$ on the randomized treatment assignment $R_i$ and possibly baseline covariates $Z_i$ for each subject $i$ in a double-blind placebo-controlled trial in the following way,

$$\mathrm{E}(Y_i - Y_{i0}|X_i, R_i = 1, Z_i) = \psi^* X_i. \tag{2.8}$$

Causal parameter $\psi^*$ is an unknown finite-dimensional parameter, which expresses the expected change in outcome when those exposed to $X_i = 1$ would have their treatment set to $X_i = 0$ (placebo). These models make no direct assumption on selected treatment compliance levels and placebo prognosis but rely on the IV assumption and a parametric form for the causal effects (Goetghebeur and Lapp, 1997). These are special cases of structural nested mean models (Robins, 1994; Goetghebeur and Lapp, 1997). In the DBP trial, exposures are the daily amount of experimental drug intake or dose as measured by electronic devices. The baseline covariates $Z$=(age, sex, height, weight, and diastolic blood pressure) are measured before randomization (pre-randomization). For each subject $i = 1, ..., n$, we assume observed data $(Y_i, X_i, R_i, Z_i)$ represent $n$ independent and identically distributed random vectors. Primary interest lies in the effect of treatment on expected diastolic blood pressure reduction from baseline (i.e. the time of active randomization). Note that, the treatment compliance level $X_i$ is a post-randomization variable and $Z_i$ contains pre-randomization variables. As stated before, because the potential treatment-free outcome $Y_{i0}$ is not observed (it is sometimes called latent treatment- free outcome), we would not be able to estimate causal parameter $\psi^*$ without some assumptions. The information necessary to estimate $\psi^*$ can be drawn from the combination of the following 3 assumptions (Robins, 1994; Goetghebeur and Lapp, 1997; Vansteelandt and Goetghebeur 2003, 2004; Goetghebeur and Vansteelandt 2005).

1. To link treatment-free outcome to observed outcome, we assume that $Y_i = Y_{i0}$ (consistency assumption) for subjects with $X_i = 0$. In fact, this assumption

states that on both arms, the observed outcome in the actually untreated subset ($X_i = 0$) corresponds to $Y_{i0}$. Under this assumption indeed, $\mathrm{E}(Y_i - Y_{i0}|X_i = 0, R_i, Z_i) = 0$.

2. Treatment-free outcomes have equal averages in both randomized arms, within strata of baseline covariates $Z_i$ (randomization assumption), $\mathrm{E}(Y_{i0}|Z_i, R_i) = \mathrm{E}(Y_{i0}|Z_i)$. This assumption holds in well-conducted blinded randomized trials because treatment-free outcomes are not affected by received treatment and can therefore be envisaged as fixed characteristics of each subject.

3. The expected causal effect (2.8) follows the LSMM (model assumption).

Under assumptions (1)-(3) and further mild regularity conditions, a consistent, asymptotically normal estimator for $\psi^*$ can be obtained for active treatment $A$ versus placebo. Under model (2.8), we derive

$$\mathrm{E}(Y_i - \psi^* X_i R_i|X_i, R_i, Z_i) = \mathrm{E}(Y_{i0}|X_i, R_i, Z_i).$$

It follows from the randomization assumption that

$$\begin{aligned} \mathrm{E}(Y_i - \psi^* X_i R_i|R_i = 1, Z_i) &= \mathrm{E}\{\mathrm{E}(Y_i - \psi^* X_i R_i|X_i, R_i = 1, Z_i)|R_i = 1, Z_i\} \\ &= \mathrm{E}(Y_{i0}|R_i = 1, Z_i) = \mathrm{E}(Y_{i0}|Z_i), \end{aligned}$$

and similarly, $\mathrm{E}(Y_i|X_i, R_i = 0, Z_i) = \mathrm{E}(Y_{i0}|R_i = 0, Z_i) = \mathrm{E}(Y_{i0}|Z_i)$, since $X_i \equiv 0$. Then

$$\mathrm{E}(Y_i - \psi^* X_i R_i|R_i = 1, Z_i) = \mathrm{E}(Y_i - \psi^* X_i R_i|R_i = 0, Z_i). \tag{2.9}$$

In words, the latter states that, after subtracting the average causal effect of $X_i$ from the observed outcome $Y_i$ on the treatment arm, both randomized groups have the same average at fixed levels of baseline covariates $Z_i$. Therefore,

$$g(Z_i)\mathrm{E}\left[\{R_i - P(R_i = 1|Z_i)\}\{Y_i - \psi^* X_i R_i - q(Z_i)\}|Z_i\right] = 0$$

for any finite-dimensional function $g(Z_i)$ and scalar function $q(Z_i)$. This implies that by solving the following unbiased estimating equation,

$$\sum_{i=1}^{n} U_i(\psi^*) = \sum_{i=1}^{n} g(Z_i)\left[\{R_i - P(R_i = 1|Z_i)\}\{Y_i - \psi^* X_i R_i - q(Z_i)\}\right] = 0 \tag{2.10}$$

we can find an estimator $\hat{\psi^*}$ of $\psi^*$. The estimating equation $U_i(\psi^*)$ is unbiased; that is, $\mathrm{E}\{U_i(\psi^*)\} = 0$ because

$$\begin{aligned}
\mathrm{E}\{U_i(\psi^*)\} &= \mathrm{E}\left[\mathrm{E}\{U_i(\psi^*)|Z_i\}\right] \\
&= \mathrm{E}[\mathrm{E}\left[\mathrm{E}\{U_i(\psi^*)|R_i, Z_i\}|Z_i\right]] \\
&= \mathrm{E}\left[\mathrm{E}[g(Z_i)\{R_i - P(R_i = 1|Z_i)\}\right. \\
&\quad \left.\times\{\mathrm{E}(Y_i - \psi^* X_i R_i|Z_i, R_i) - q(Z_i)\}|Z_i]\right] \\
&= \mathrm{E}\left[g(Z_i)\mathrm{E}\{R_i - P(R_i = 1|Z_i)|Z_i\}\mathrm{E}\{Y_i - \psi^* X_i R_i - q(Z_i)\}|Z_i\}\right] \\
&= 0.
\end{aligned}$$

Therefore, the resulting estimator $\hat{\psi^*}$ is a consistent and asymptotically normal estimator for $\psi^*$ under model (2.8) (Robins, 1994; Goetghebeur and Lapp, 1997). For each choice of $g(Z_i)$ and $q(Z_i) = \mathrm{E}(Y_{i0}|Z_i)$, the solution to (2.10) yields a consistent estimator under assumptions (1)-(3) with known $P(R_i = 1|Z_i)$. Note that, an optimal choice of $g(Z_i)$ and $q(Z_i)$ will yield the semi-parametrically efficient estimate for $\psi^*$. Note also that $\hat{\psi^*}$ can be made more efficient as $Z_i$ becomes more predictive of compliance. "When $\hat{\psi^*}$ is found, it can be used to calculate expected treatment-free outcomes $Y_{i0} = Y_i - \hat{\psi^*} X_i R_i$ on both arms and to check whether their regression on $X_i$ coincides on both arms as dictated by (2.9)". Under model (2.8), a closed form solution $\psi^*$ can be obtained as a function of $g(Z_i)$ and $q(Z_i)$:

$$\hat{\psi^*} = \frac{\sum_{i=1}^{n} g(Z_i)\{R_i - P(R_i = 1|Z_i)\}\{Y_i - q(Z_i)\}}{\sum_{i=1}^{n} g(Z_i)\{R_i - P(R_i = 1|Z_i)\}X_i}.$$

Following Robins (1994) and Goetghebeur and Lapp (1997), a consistent estimator for the variance of $\hat{\psi^*}$ can also be derived from

$$Var(\hat{\psi^*}) = \hat{\Omega}/\hat{\tau}^2$$

where $\hat{\Omega} = 1/n \sum_{i=1}^{n} U_i(\hat{\psi^*}) U_i^t(\hat{\psi^*})$ and $\hat{\tau} = 1/n \sum_{i=1}^{n} g(Z_i)\{P(R_i = 1|Z_i) - R_i\}X_i R_i$. For instance, in the DBP trial data with $q(Z_i) = \mathrm{E}(Y_i - \psi^* R_i X_i|Z_i)$ and $g(Z_i) = Z_i$ where $Z_i$ contains all baseline covariates, we obtain $\hat{\psi^*} = -7.51$ with 95 % confidence interval $(-11.09, -3.93)$.

## 6.3 Error manifest and estimation process

As stated, in spite of recent technically sophisticated tools, for measuring compliance with an assigned drug dosing regimen, errors in exposure measurement are

inevitable. As discussed in chapter 1, when actual exposure level $X_i$ is imperfectly measured, we observe $W_i$ instead of $X_i$ which may potentially differ from $X_i$ for each subject $i = 1, ..., n$. Therefore identification of the causal effect under model (2.8) requires additional assumptions that we will investigate in chapter 4. In this section we illustrate how, for instance, the estimation approach under (2.8) will proceed with the measurement error problem. As discussed in chapter 1, for our example, suppose that the error on exposure $X_i$ follows the classical measurement error model (a similar development can be made for the Berkson error model), that is, $W_i = X_i + U_i$ with $U_i \perp\!\!\!\perp (X_i, Z_i)$, where $U_i$ is a measurement error term. In the context of measurement error, it is often assumed that the error term $U_i$ has mean zero with homoscedastic variance $\sigma_u^2$ which is known or can be estimated from an extra data set. Here, we assume that $\mathrm{E}(W_i - X_i|Z_i, R_i) = \mathrm{E}(U_i|Z_i, R_i = 1) = \delta^*$; that is $U_i$ has conditional mean $\delta^*$ which is the same quantity as in the classical measurement error model. This indeed considers average measurement error in groups determined by levels of covariate $Z_i$ to be constant, a restriction that can be relaxed. When $\delta^* = 0$ for all $Z_i$, that is $\mathrm{E}(W_i|Z_i, R_i = 1) = \mathrm{E}(X_i|Z_i, R_i = 1)$, we say that the measurement error is conditionally unbiased. Note that we are using the notation $U$ for unmeasured confounders, $U_i(.)$ for estimating equations, and here $U_i$ as a measurement error term. Now our observed data is $(Y_i, W_i, R_i, Z_i)$. We can write the conditional mean independence (2.9) for the observed data as

$$\mathrm{E}\{Y_i - \psi^*(W_i - \delta^*)R_i|R_i, Z_i\} = \mathrm{E}\{Y_i - \psi^*(W_i - \delta^*)R_i|Z_i\}.$$

Note that, this is designed to make the predicted treatment-free outcome, $Y_i - \psi^*(W_i - \delta^*)R_i$ mean independent of $R_i$ conditional on $Z_i$. By this, under model (2.8) and an uncontaminated control group, we show that the following estimating equation is unbiased with arbitrary index functions $g(Z_i)$ and $q(Z_i)$

$$\sum_{i=1}^{n} U_i(\psi^*, \delta^*) = \sum_{i=1}^{n} g(Z_i)\{R_i - P(R_i = 1|Z_i)\}\{Y_i - \psi^*(W_i - \delta^*)R_i - q(Z_i)\} = 0$$

$$\begin{aligned}
\mathrm{E}\{U_i(\psi^*, \delta^*)\} &= \mathrm{E}[\mathrm{E}\left[\mathrm{E}\{U_i(\psi^*, \delta^*)|R_i, Z_i\}|Z_i\right]] \\
&= \mathrm{E}[\mathrm{E}[g(Z_i)\{R_i - P(R_i = 1|Z_i)\} \\
&\quad \times\{\mathrm{E}(Y_i - \psi^*(W_i - \delta^*)R_i|Z_i, R_i) - q(Z_i)\}|Z_i]] \\
&= \mathrm{E}\left[g(Z_i)\mathrm{E}\{R_i - P(R_i = 1|Z_i)|Z_i\}\right. \\
&\quad \left.\times\mathrm{E}\{Y_i - \psi^*(W_i - \delta^*)R_i - q(Z_i)\}|Z_i\}\right] \\
&= 0.
\end{aligned}$$

Following Goetghebeur and Vansteelandt (2005), and Vansteelandt and Goetghebeur (2003), the most efficient estimation in the class determined by the latter unbiased estimating equation can be obtained from $q_{opt}(Z_i) = \mathrm{E}(Y_i|Z_i, R_i = 0)$ and

$$g_{opt}(Z_i) = \{\mathrm{E}(W_i|Z_i, R_i = 1) - \delta^*\}Z_i \times [\{1 - P(R_i = 1|Z_i)\}Var\{Y_i - \psi^*(W_i - \delta^*)Z_i|Z_i, R_i = 1\} + P(R_i = 1|Z_i)Var(Y_i|Z_i, R_i = 0)]^{-1}.$$

## 6.4 The impact of biased measurement error

We examine the potential impact of biased measurement error on estimated treatment effects under model (2.8) in the context of the DBP trial. For doing so, we reanalyze both active treatments $A$ and $B$ separately versus placebo assuming the over reporting of dose is for instance 0 %, 10 % or 20 %. Specifically, we reanalyze the data, assuming $\mathrm{E}(W - X|Z, R) = \mathrm{E}(U|Z, R) = \delta^*$ is equal to, 0 %, 10 % or 20 % with and without adjustment for baseline covariates, where 0 % encodes the situation where measurement error is ignored. This considers that when not all

Table 2.6: *Sensitivity analysis for the DBP trial under LSMM with and without adjustment for covariates, when the average measurement error is 0 % , 10 % or 20 %. The first row refers to $Z = 1$ (no covariates) and the second row is adjusting for covariates. Treatment A refers to 'Nebivolol' and treatment B refers to 'Atenolol'.*

| | $\delta = 0.0$ | $\delta = 0.10$ | | $\delta = 0.20$ | |
|---|---|---|---|---|---|
| | Classical | Classical | Berkson | Classical | Berkson |
| A vs. Plac. | 9.56 (2.10) | 10.84 (2.39) | 8.55 (1.87) | 12.51 (2.79) | 7.73 (1.68) |
| B vs. Plac. | 10.37 (2.41) | 11.64 (2.72) | 9.36 (2.17) | 13.25 (3.12) | 8.52 (1.97) |
| A vs. Plac. | 7.51 (1.83) | 8.50 (2.06) | 6.73 (1.64) | 9.77 (2.38) | 6.09 (1.48) |
| B vs. Plac. | 7.50 (2.18) | 8.37 (2.45) | 6.79 (1.96) | 9.47 (2.79) | 6.20 (1.78) |

openings of the pill container correspond to the actual pill intake, exposure will be over reported. This sensitivity analysis has been done for active treatment $A$ versus placebo by Goetghebeur and Vansteelandt (2005). Table 2.6 summarizes the results when the error in active treatments follows the classical error model or the Berkson error model. Note that when $\delta^* = 0$, the estimated average blood pressure reduction with the Classical error model and the Berkson error model are the same. Table 2.6

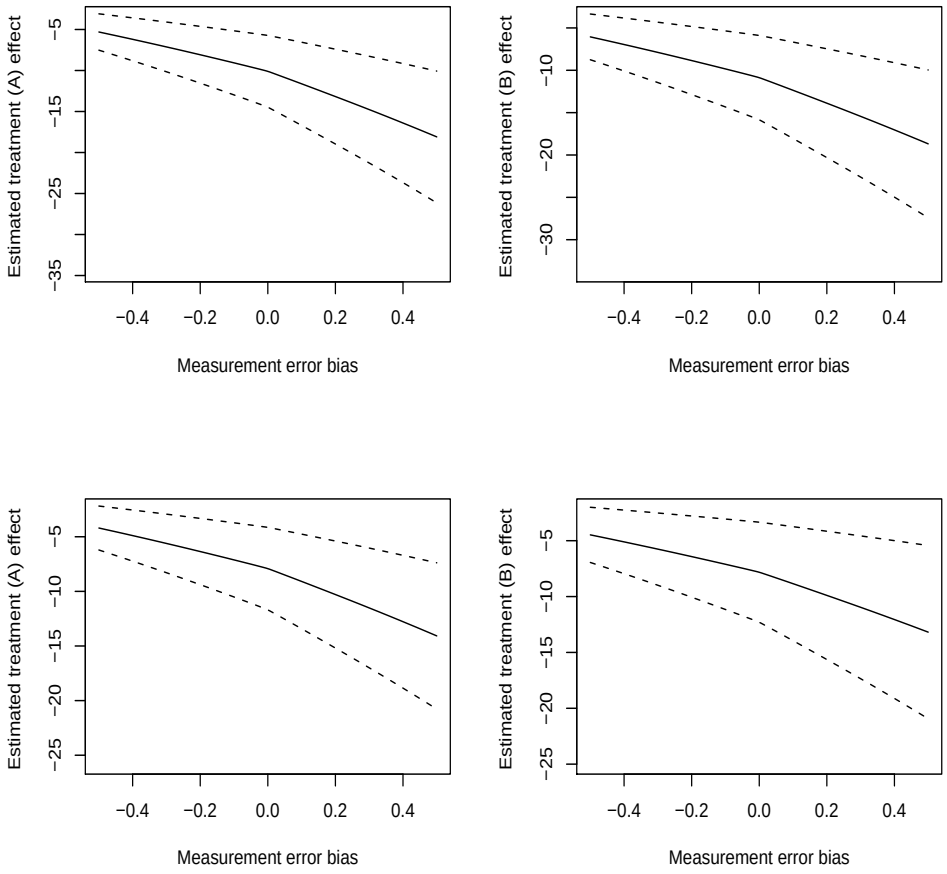



Figure 6: *Estimated average blood pressure $\hat{\psi}^*$ with 95 % confidence interval in function of $\delta^*$ for both treatments A='Nebivolol' and B='Atenolol' separately versus placebo for the DBP trial. The first row refers to $Z = 1$ (no covariates) and the second row is adjusting for covariates. Dotted lines show the upper and lower bound of confidence intervals.*

shows the estimated average blood pressure reduction with standard error as a sensitivity analysis for the DBP trial. The second and third columns of Table 2.6 show that when the average measurement error $\delta^*$=10 % or 20 %, the analysis under the LSMM estimates a more substantial blood pressure decrease for a given treatment dose under the classical error model than under the Berkson error. Standard errors are however not small. As a result, this analysis shows how ignoring measurement error will yield biased estimates and misleading inference. In Figure 6, we let the constant average error vary over the interval (-0.5, 0.5) and display the impact of biased measurement error on estimated treatment effects in model (2.8). The top row refers to $Z = 1$ (no covariates) and the second row is adjusting for covariates.

## 6.5 Log-linear and logistic structural mean models

When the outcome of interest is strictly positive, one may postulate a log-linear structural mean model (Robins, 1994, 1997; Vansteelandt and Goetghebeur, 2003) which is a special case of the multiplicative structural mean model (SMM). This model in fact assumes that a linear function of dose, possibly adjusted for covariates, explains the difference in log means of treated and potential treatment-free outcomes. For instance, the model for a constant by linear treatment effect is:

$$\log\{\mathrm{E}(Y_i|X_i, Z_i, R_i = 1)\} - \log\{\mathrm{E}(Y_{i0}|X_i, Z_i, R_i = 1)\} = \psi^* X_i. \tag{2.11}$$

This model "implies that if some group on the treatment arm received the zero dose, $X = 0$, its expected values of the observed and potential treatment-free outcomes would coincide". With a dichotomous outcome, model (2.11) indicates that the probability of success for subjects who received treatment $X_i = 1$ would become $\exp(\psi^*)$ smaller if they had not received the treatment . These parameters describe different aspects of the causal risk ratio

$$\frac{P(Y_i = 1|X_i, Z_i, R_i = 1)}{P(Y_{i0} = 1|X_i, Z_i, R_i = 1)} = \exp(\psi^* X_i).$$

It follows under this model that

$$\mathrm{E}(Y_i|X_i, Z_i, R_i = 1)\exp(-\psi^* X_i) = \mathrm{E}(Y_{i0}|X_i, Z_i, R_i = 1).$$

Key to the identification of $\psi^*$ is the randomization assumption, $Y_{i0} \perp\!\!\!\perp R_i|Z_i$, which implies

$$\mathrm{E}\{Y_i \exp(-\psi^* X_i)|Z_i, R_i\} = \mathrm{E}\{Y_i \exp(-\psi^* X_i)|Z_i\}.$$

This expression can be obtained by virtue of the randomization assumption

$$\begin{aligned}\mathrm{E}\{Y_i\exp(-\psi^* X_i)|Z_i, R_i=1\} &= \mathrm{E}\left[\mathrm{E}\{Y_i\exp(-\psi^* X_i)|X_i, Z_i, R_i=1\}|Z_i, R_i=1\right]\\ &= \mathrm{E}\{\mathrm{E}(Y_{i0}|X_i, Z_i, R_i=1)|Z_i, R_i=1\}\\ &= \mathrm{E}(Y_{i0}|Z_i, R_i=1)=\mathrm{E}(Y_{i0}|Z_i),\end{aligned}$$

and similarly, $\mathrm{E}\{Y_i\exp(-\psi^* X_i)|Z_i, R_i=0\} = \mathrm{E}(Y_{i0}|Z_i)$. The set of unbiased estimating equations can be obtained by replacing $Y_i - \psi^* X_i R_i$ by $Y_i\exp(-\psi^* X_i)$ in (2.10). Note that, when outcome is dichotomous, there is a closed form solution for estimating $\psi^*$.

Researchers were also interested in knowing whether actual exposure would increase the probability of success $Y_i = 1$. In a classical regression model, we model the causal effect of exposure level $X_i = x$ on the probability of success via a logistic regression model:

$$\mathrm{logit}P(Y_i=1|X_i=x)=\beta_0^*+\beta_1^* x.$$

This model implies that $\frac{P(Y=1|X=x)/P(Y=0|X=x)}{P(Y=1|X=0)/P(Y=0|X=0)}=\exp(\beta_1^* x)$. We can similarly model the causal effect of exposure on outcome via the logistic structural mean model. Models (2.8) and (2.11) are natural models for real and positive outcomes respectively, but they may fail to yield probabilities of success between 0 and 1 when accommodating dichotomous data (Vansteelandt and Goetghebeur, 2003). For this reason, one may consider the logistic structural mean models instead (Vansteelandt and Goetghebeur, 2003; Robins and Rotnitzky, 2004)

$$\mathrm{logit}\{P(Y_i=1|X_i, Z_i, R_i=1)\}-\log\{P(Y_{i0}=1|X_i, Z_i, R_i=1)\}=\psi^* X_i. \quad (2.12)$$

Here, the unknown causal parameter $\psi^*$ compares the success odds under observed compliance on the treatment arm with the success odds for the same group of people had they received placebo. In fact, their parameters describe different aspects of the causal odds ratio

$$\frac{P(Y_i=1|X_i, Z_i, R_i=1)/P(Y_i=0|X_i, Z_i, R_i=1)}{P(Y_{i0}=1|X_i, Z_i, R_i=1)/P(Y_{i0}=0|X_i, Z_i, R_i=1)}=\exp(\psi^* X_i).$$

Although the randomization assumption still holds here, the problem of the logistic SMMs stems from the fact that, it is hard to predict:

$$\begin{aligned}\mathrm{E}(Y_{i0}|Z_i, R_i=1) &= \int \mathrm{expit}\left[\mathrm{logit}\{P(Y_i=1|X_i, Z_i, R_i=1)\}\right.\\ &\quad \left.-\psi^* X_i\right] f(x_i|Z_i, R_i=1)dx_i\end{aligned}$$

where $\text{expit}(a) = \exp(a)/\{1 + \exp(a)\}$. Therefore, to obtain an estimating equation the estimation approach outlined for linear and log-linear SMMs, does not work for logistic SMMs. Robins (1999) showed that $\psi^*$ in model (2.12) can not be estimated with G-estimation. Vansteelandt and Goetghebeur (2003) solve this problem by proposing IVs estimators for $\psi^*$ under an additional model for the observed relationship between outcome and exposure in the experimental arm. They postulate for instance a logistic regression model

$$\begin{aligned} \text{logit}P(Y_i = 1|X_i, Z_i, R_i = 1) &= m(X_i, Z_i, R_i = 1; \beta^*) \\ e.g. &= \beta_0^* + \beta_1^* X_i + \beta_1^* Z_i + \beta_1^* X_i Z_i \end{aligned}$$

where $\beta^*$ can be estimated by logistic regression model. Under the randomization assumption and some further mild regularity conditions, they find the unbiased estimating equation

$$\sum_{i=1}^{n} g(Z_i)\{R_i - P(R_i = 1|Z_i)\}[R_i \text{expit}\{m(X_i, Z_i, R_i = 1; \beta^*) - \psi^* X)\} + (1 - R_i)Y_i] = 0,$$

and by solving this, a consistent estimate for $\psi^*$ under model (2.12) can be obtained.

# Chapter 3

# On the Performance of Instrumental Variable Estimators of the Causal Odds Ratio


## Summary

Inference for causal effects can benefit greatly from the availability of an instrumental variable (IV) which, by definition, is associated with the given exposure, but not with the outcome of interest except through a causal exposure effect. Estimation methods for instrumental variables are now well established for continuous outcomes. The case of dichotomous outcomes turns out much more difficult and has received far less attention to date. In this article, we give an expository review of exact as well as approximate IV-estimators for the causal odds ratio, that have been proposed in the biostatistical, epidemiological and econometric literature. Methods comparisons are made, both theoretically and via extensive simulation, and new insights are developed into the assumptions underlying their validity. The different estimators are used to assess the risk of gastrointestinal (GI) complications attributable to different non-steroidal anti-inflammatory drugs (instead of Cox-2 inhibitors).

# 1 Introduction

It is well known that most causal analyses of observational data rely heavily on the untestable assumption of no unmeasured confounders. According to this assumption, one has available all prognostic factors of the exposure that are *also* associated with the outcome *other than via* a possible exposure effect on outcome. Concerns about the validity of this assumption plague observational data analyses (see e.g. Prentice, Pettinger and Anderson, 2005) and increase the uncertainty surrounding many study results (Greenland, 2005; Vansteelandt et al., 2006). This is especially true in settings where the data analysis is based on registry data or focuses on research questions different from those conceived at the time of data collection. Substantial progress can be made in settings where measurements are available on a so-called instrumental variable (IV). This is a prognostic factor of the exposure, which is *not* associated with the outcome, *except via* a possible exposure effect on outcome (Hernán and Robins, 2006).

IVs have a long tradition in econometrics and are becoming increasingly popular in biostatistics and epidemiology. This is partly because the existence of an IV is sometimes guaranteed by design. For instance, randomized encouragement designs whereby, say, smoking pregnant women are randomly assigned to intensified encouragement to quit smoking or not, yield - by virtue of randomization - a valid IV for assessing the effects of smoking on low birth weight (Permutt and Hebel, 1989). The growing success of IV methods in biostatistics and epidemiology can, however, be mainly attributed to applications in genetic epidemiology (Smith and Ebrahim, 2004). Here, the random assortment of genes transferred from parents to offspring resembles the use of randomization in experiments and is therefore often referred to as 'Mendelian randomization'. Building on this idea, genetic variants may sometimes qualify as an IV for estimating the relationship between a genetically affected exposure and a disease outcome (Didelez and Sheehan, 2007; Lawlor et al., 2008).

Estimation methods for IVs are now well established for continuous outcomes. The case of dichotomous outcomes turns out much more difficult and has received far less attention to date. This paper therefore combines different, scattered developments in the biostatistical, epidemiological and econometric literature and aims to improve the clarity and comparability of these developments by casting them within a common causal framework based on counterfactuals. This will yield new insights, in particular into the assumptions underlying each of the considered methods and into their successfulness at approximating the causal odds ratio, which we define explicitly in the next section. In Sections 2.2 and 2.3, we focus on exact IV methods for the causal odds ratio under logistic structural mean models, as proposed by Vansteelandt and Goetghebeur (2003, 2005) and later extended in Robins and Rotnitzky (2004). In

particular, we clarify the connections between both approaches. We then consider a number of approximate IV estimators that are popular in epidemiology (see Sections 2.3 and 2.4) and in econometrics (see Section 2.5). Both theoretical arguments and extensive simulation are used to contrast the different estimators, which are eventually used to assess the risk of gastrointestinal (GI) complications attributable to different non-steroidal anti-inflammatory drugs (instead of Cox-2 inhibitors). Because many commonly employed IV analyses do not require covariate adjustment, this will also be the motivating setting for most of the article. In the discussion section, we elaborate on covariate-adjusted IV estimators and extensions to longitudinal data.

# 2 IV-estimators of the causal odds ratio

Our goal in this article is to assess the causal effect of an arbitrary exposure $X_i$, measured for subjects $i = 1, ..., n$, on a dichotomous outcome $Y_i$. For instance, in the data analysis section 3, we will estimate the effect of Cox-2 treatment $X_i = 1$ (versus non-selective NSAIDs $X_i = 0$) on gastrointestinal bleeding (i.e. $Y_i = 1$; $Y_i = 0$ otherwise). To succeed, we will assume in this paper that we have measured for each subject $i$ an instrumental variable (IV) $Z_i$ for the effect of $X_i$ on $Y_i$ which, by definition, satisfies the following properties: (a) $Z_i$ is associated with $X_i$; (b) $Z_i$ affects the outcome $Y_i$ only through $X_i$ (i.e. often referred to as the exclusion restriction); (c) the association between $Z_i$ and $Y_i$ is unconfounded (i.e. often referred to as the randomization assumption) (Hernán and Robins, 2006). In Section 3, as in Brookhart and Schneeweiss (2007), we choose the physician's prescribing preference for Cox-2 (versus non-selective NSAIDs) as an instrumental variable (i.e. $Z_i$) for the effect of Cox-2 treatment on gastrointestinal bleeding. This qualifies as a possible IV because it is associated with Cox-2 treatment (i.e. (a)), because it seems reasonable that the physician's prescribing preference can only affect a patient's gastrointestinal bleeding through his/her prescription (i.e. (b)) and is not otherwise associated with patient's gastrointestinal bleeding (i.e. (c)). The latter assumption would fail if for instance patients with high risk of bleeding are more often seen with physicians who prefer Cox-2 (Hernán and Robins, 2006).

## 2.1 Causal odds ratio and logistic structural mean models

We define causal effects in terms of comparisons of each subject $i$'s observed outcome with a counterfactual outcome $Y_{i0}$, which denotes the outcome value that we would have observed for that subject if the exposure were controlled at some chosen reference exposure level 0 (e.g. non-selective NSAIDs). In particular, we will

summarize causal effects in terms of the causal odds ratio

$$\frac{\text{odds}\left(Y_i = 1|X_i, Z_i\right)}{\text{odds}\left(Y_{i0} = 1|X_i, Z_i\right)}, \tag{3.1}$$

where for any $Y_i^*$, $\text{odds}\left(Y_i^* = 1|X_i, Z_i\right) \equiv P\left(Y_i^* = 1|X_i, Z_i\right)/P\left(Y_i^* = 0|X_i, Z_i\right)$. This expresses how much the odds of 'success' would change for subjects with exposure level $X_i$ and instrumental variable $Z_i$, if their exposure were set to the reference level 0. It thus measures the effect of *received* treatment, or, the so-called treatment effect *in the treated* (Hernán and Robins, 2006; Robins, VanderWeele and Richardson, 2006; Didelez, Meng and Sheehan, 2008). Because of identifiability constraints (Vansteelandt and Goetghebeur, 2005), we will assume that the causal odds ratio obeys the simple loglinear model restriction, $(3.1) = \exp(\psi^* X_i)$, where $\psi^*$ is an unknown parameter. Equivalently, we will conduct inference under the logistic structural mean model (Vansteelandt and Goetghebeur, 2003)

$$\text{logitE}(Y_i|X_i, Z_i) - \text{logitE}(Y_{i0}|X_i, Z_i) = \psi^* X_i, \tag{3.2}$$

where $\text{logit}(p) = \log\{p/(1-p)\}$. In Section 2.2, we will review exact IV methods for the causal odds ratio $\exp(\psi^* X_i)$ under model (3.2), that have been introduced in the statistical literature. In Sections 2.3 and 2.4, we will review approximate IV methods for the causal odds ratio that have been introduced in the epidemiological literature on Mendelian randomization and propose related novel procedures. In Section 2.5, we discuss generalized method of moments (GMM) estimators that are frequently considered in econometrics.

## 2.2 Exact estimation in logistic structural mean models

Although $Y_i$ may well depend on $Z_i$ (in the presence of an exposure effect), the IV-assumptions imply that $Y_{i0} \perp\!\!\!\perp Z_i$. To make use of this, Vansteelandt and Goetghebeur (2003) average over the observed exposure values in model (3.2). Because this is not possible without making additional parametric modelling assumptions (Robins and Rotnitzky, 2004), they model the expected observed outcome, conditional on the exposure and instrumental variable. For instance, one may choose

$$\text{logit}P(Y_i = 1|X_i, Z_i) = \beta_0^* + \beta_1^* X_i + \beta_2^* Z_i + \beta_3^* X_i Z_i, \tag{3.3}$$

where $\beta_0^*, \beta_1^*, \beta_2^*, \beta_3^*$ are unknown scalar parameters. More generally, one may postulate that

$$\text{logit}E(Y_i|X_i, Z_i) = m(X_i, Z_i; \beta^*), \tag{3.4}$$

where $m(X_i, Z_i; \beta)$ is a known function, smooth in $\beta$, and $\beta^*$ is an unknown finite-dimensional parameter. An estimate $\hat{\beta}$ of $\beta^*$ can be obtained using standard methods (e.g. using maximum likelihood estimation). Combining the causal model (3.2) with the so-called association model (3.4) yields a prediction for the counterfactual outcome $Y_{i0}$ for each subject $i$ which, for given $\psi$, equals

$$H_i(\psi, \hat{\beta}) = \text{expit}\{m(X_i, Z_i; \hat{\beta}) - \psi X_i\},$$

where $\text{expit}(a) \equiv \exp(a)/\{1 + \exp(a)\}$. Because

$$\text{E}(Y_{i0}|Z_i) = \text{E}(Y_{i0})$$

under the IV-assumptions, the value of $\psi^*$ can now be chosen as the value $\psi$ which makes the empirical means equal, once $Y_{i0}$ is replaced by $H_i(\psi, \hat{\beta})$. For a dichotomous instrument $Z_i$, taking the values 0 and 1, one thus chooses $\psi$ such that

$$\frac{\sum_i H_i(\psi, \hat{\beta}) Z_i}{\sum_i Z_i} = \frac{\sum_i H_i(\psi, \hat{\beta})(1 - Z_i)}{\sum_i (1 - Z_i)}. \tag{3.5}$$

When also the exposure is dichotomous, model (3.3) is guaranteed to hold and the following closed-form estimator is obtained:

$$\hat{\psi} = \log \left[ \frac{-Q_1 \pm \sqrt{Q_1^2 - 4Q_2(Q_2 - \hat{X}_{11} + \hat{X}_{10})Q_3}}{2Q_2} \right], \tag{3.6}$$

where $\hat{X}_{xz}$ is the percentage of subjects with $X = x$ amongst those with $Z = z$, and

$$\begin{aligned} Q_1 &= (Q_2 + \hat{X}_{10}) \exp(\hat{\beta}_0 + \hat{\beta}_1) + (Q_2 - \hat{X}_{11}) \exp(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3) \\ Q_2 &= \text{expit}(\hat{\beta}_0)\hat{X}_{00} - \text{expit}(\hat{\beta}_0 + \hat{\beta}_2)\hat{X}_{01} \\ Q_3 &= \exp(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3) + \exp(\hat{\beta}_0 + \hat{\beta}_1). \end{aligned}$$

In most cases, (3.6) yields a unique estimator of the causal odds ratio, although multiple or no solutions are rarely obtained when precision is limited due to small sample size or outcome mean close to 0 or 1. This is illustrated in Figure 1, which displays the left- and righthand side of (3.5) in function of $\psi$ for 3 settings. The top 2 panels are based on the same simulated data set. They show that 2 or no solutions can be obtained for the same data set, depending on whether the association model (3.4) includes an interaction between exposure and instrument (left panel) or not (right panel). The bottom panel corresponds to the data analysis of Section 3, where

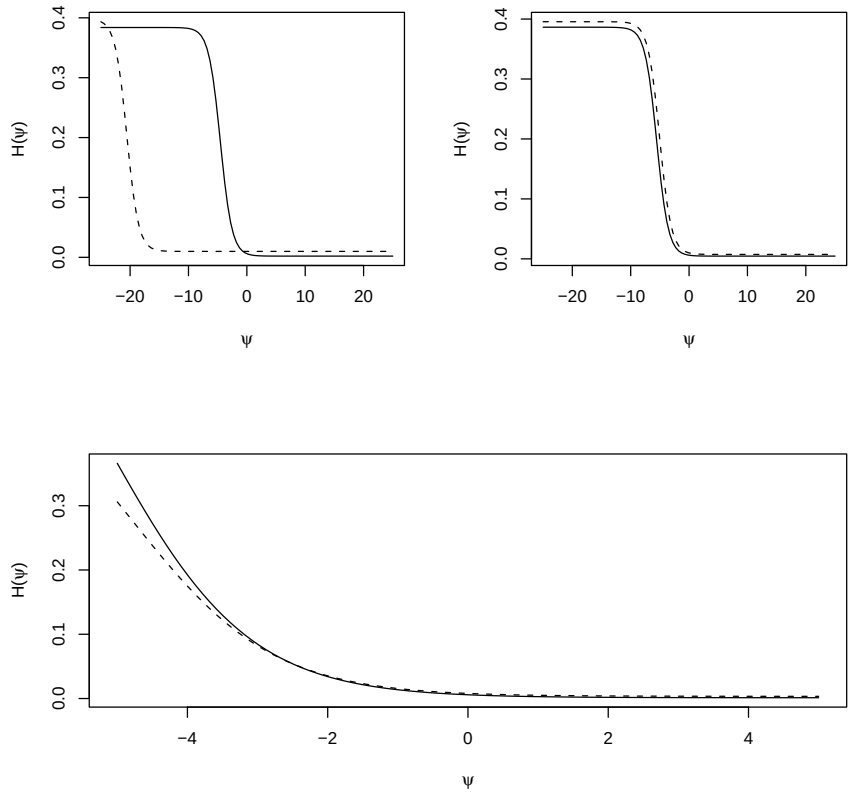



Figure 1: *Plot of the left- (solid) and righthand side (dotted) of expression (3.5) as a function of $\psi$. Top: simulated data set (Right: with $\beta_4^* = 0$ in model (3.3)); Bottom: observed data set.*

a single solution was obtained. Our experience indicates that, when 2 solutions are obtained, one of them usually corresponds to an effect size which is so large that it would be deemed unrealistic. When no solutions are obtained, this can sometimes be resolved by choosing a less parsimonious association model (as in Figure 1, top) or by adjusting for covariates (see the discussion section). For general instruments, a consistent point estimator of $\psi^*$ can be found by solving unbiased estimating equation

$$0 = \sum_{i=1}^{n} \left[ d(Z_i) - \frac{\sum_{j=1}^{n} d(Z_j)}{n} \right] H_i(\psi, \hat{\beta}) \tag{3.7}$$

for $\psi$, where $d(Z_i)$ is an arbitrary function of $Z_i$, e.g. $d(Z_i) = Z_i$ (see Vansteelandt and Goetghebeur (2003) for choices that yield a semiparametric efficient estimator of $\psi^*$). This thus leads to the following 2-stage approach:

1. First fit the association model (3.4), for instance using maximum likelihood estimation, and obtain an estimate $\hat{\beta}$ of $\beta^*$;

2. Next, solve equation (3.7) to obtain an estimate $\hat{\psi}$ of $\psi^*$.

The resulting estimator will be referred to as 'Exact IV-estimator I' throughout. In the Appendix, we show that when the association model includes an intercept and main effect in $Z_i$ (as in model (3.3)) and is fitted using maximum likelihood estimation in standard generalized linear model software, then its solution is robust to misspecification of the association model (3.4) when $\psi^* = 0$. This approach thus yields a valid test of the causal null hypothesis that $\psi^* = 0$, even when both models (3.2) and (3.4) are misspecified. This 'local' robustness property (Vansteelandt and Goetghebeur, 2003) also guarantees that estimators of the causal odds ratio will have small bias under model misspecification when the true exposure effect is close to, but not equal to zero.

A drawback of the parameterization by Vansteelandt and Goetghebeur (2003) is that the association model may be incongenial with the causal model. Specifically, there may be no values of the causal parameter $\psi$ for which $\mathrm{E}\{H_i(\psi, \beta^*)|Z_i\} = \mathrm{E}\{H_i(\psi, \beta^*)\}$ over the entire support of $Z_i$ under the model (where $\beta^*$ now corresponds to the limiting value of $\hat{\beta}$). In the Appendix, we show that this may happen when parametric restrictions are imposed on the main effect of the instrumental variable in the association model (3.4), but not when that main effect is left unrestricted. In the common situation of a dichotomous instrument, this imposes no limitations on the applicability of this method so long as a main effect of the IV is included in the association model. When the instrument is categorical with 3 levels, as is often the

case in Mendelian randomization studies where a marker coding is used as the instrument, this requires using two dummy regressors for the instrument in the association model. For general IVs, this would require using a generalized additive association model which leaves the main effect of the IV unrestricted (apart from smoothness restrictions).

Robins and Rotnitzky (2004) developed an alternative approach for estimation of $\psi^*$ in model (3.2), which guarantees a congenial parameterization by avoiding direct specification of an association model. They parameterize instead the selection-bias function

$$\text{logitE}(Y_{i0}|X_i, Z_i) - \text{logitE}(Y_{i0}|X_i = 0, Z_i) = q(X_i, Z_i; \eta^*) \tag{3.8}$$

where $q(X_i, Z_i; \eta)$ is a known function satisfying $q(0, Z_i; \eta) = 0$, smooth in $\eta$, and $\eta^*$ is an unknown finite-dimensional parameter. That $q(X_i, Z_i; \eta^*)$ encodes the degree of selection bias can be seen because $q(X_i, Z_i; \eta^*) = 0$ for all $X_i$ implies that $\text{E}(Y_{i0}|X_i, Z_i) = \text{E}(Y_{i0}|Z_i)$ and thus implies that the association between exposure and outcome (more precisely, $Y_{i0}$) is unconfounded (conditional on $Z_i$). The approach of Robins and Rotnitzky (2004) relies on the following iterative procedure (which is here simplified because of the absence of covariate adjustment):

1. Compute a maximum likelihood estimator $\hat{\alpha}$ for the finite-dimensional parameter $\alpha^*$ indexing a model for the conditional exposure density $P(X_i|Z_i; \alpha^*)$;

2. For fixed $\psi$ (starting from an initial value $\psi_0$), compute maximum likelihood estimators $\hat{\eta}(\psi)$ and $\hat{\omega}(\psi)$ for the parameters $\eta^*$ and $\omega^*$ indexing the implied association model

$$P(Y_i = 1|X_i, Z_i; \psi, \eta^*, \omega^*) = \text{expit}\{\psi X_i + q(X_i, Z_i; \eta^*) + v(Z_i; \eta^*, \omega^*)\} \tag{3.9}$$

where $v(Z_i; \eta^*, \omega^*) \equiv \text{logit}\{E(Y_{i0}|X_i = 0, Z_i)\}$ must satisfy

$$\begin{aligned} \omega^* &\equiv \text{logitE}(Y_{i0}) \\ &= \int \text{expit}\{q(X_i = x, Z_i; \eta^*) + v(Z_i; \eta^*, \omega^*)\} P(X_i = x|Z_i; \alpha^*) dx \end{aligned} \tag{3.10}$$

3. Solve the following estimating equation for $\psi$:

$$0 = \sum_{i=1}^{n} \left[ d(Z_i) - \frac{\sum_{j=1}^{n} d(Z_j)}{n} \right] H_i(\psi, \hat{\alpha}, \hat{\eta}(\psi), \hat{\omega}(\psi)) \tag{3.11}$$

where

$$\begin{aligned} H_i(\psi,\alpha,\eta,\omega) &= \frac{\text{expit}(M_{2i})\{1-\text{expit}(M_{2i})\}}{\text{expit}(M_{1i})\{1-\text{expit}(M_{1i})\}}\{Y_i-\text{expit}(M_{1i})\} \\ &\quad +\text{expit}(M_{2i}) \\ M_{1i} &= \psi X_i + q(X_i,Z_i;\hat{\eta}(\psi)) + v(Z_i;\hat{\eta}(\psi),\hat{\omega}(\psi)) \\ M_{2i} &= q(X_i,Z_i;\hat{\eta}(\psi)) + v(Z_i;\hat{\eta}(\psi),\hat{\omega}(\psi)) \end{aligned}$$

and where the choice of $d(Z_i)$ is specified next. Note the similarity between estimating equations (3.7) and (3.11): the first contribution of $H_i(\psi,\alpha,\eta,\omega)$ is a scaled residual of the implied association model (3.9), which yields an approximately zero contribution to the estimating equation, regardless of the value of $\psi$; the second contribution, $\text{expit}(M_{2i})$, corresponds to $H_i(\psi,\hat{\beta})$ in (3.7).

4. Repeat steps 2 and 3, updating $\psi$ in step 2 with the estimate obtained in step 3, until convergence.

The resulting estimator of $\psi^*$ will be referred to as 'Exact IV-estimator II' throughout. It is consistent and asymptotically normal when, besides the logistic structural mean model (3.2), model (3.8) for the selection bias function and the model for the exposure distribution $P(X_i|Z_i;\alpha^*)$ are all correctly specified. At the null hypothesis that $\psi^* = 0$, the estimator is (locally) robust against misspecification of these models.

When the instrument is dichotomous, $d(Z_i) = Z_i$ is the only possible function of $Z_i$, up to linear transformations. When the instrument is continuous or discrete with more than two levels, a semiparametric efficient estimator of $\psi^*$ can be obtained by choosing

$$\begin{aligned} d(Z_i) &= d_2(Z_i) - \frac{\text{E}\{d_2(Z_i)\}}{\text{E}\{d_3(Z_i)\}}d_3(Z_i) \\ d_2(Z_i) &= d_3(Z_i)\left(\text{E}\left\{\text{expit}(M_{2i})\{1-\text{expit}(M_{2i})\}X_i|Z_i\right\}\right. \\ &\quad \left.-\text{E}\{\text{expit}(M_{2i})\{1-\text{expit}(M_{2i})\}X_i\}\right) \\ d_3(Z_i) &= \left(\text{E}\left[\frac{\{\text{expit}(M_{2i})\{1-\text{expit}(M_{2i})\}\}^2}{\text{expit}(M_{1i})\{1-\text{expit}(M_{1i})\}}|Z_i\right] + \text{var}\{\text{expit}(M_{2i})|Z_i\}\right)^{-1}. \end{aligned}$$

In both cases, however, we recommend Exact IV-estimator I for practical data analysis (a) because it is equally valid for discrete instruments and approximately so for continuous instruments when smoothing techniques are used to model their association with outcome; (b) because it is computationally much simpler as it does not involve iteratively solving estimating equations, nor solving integral equation (3.11); and (c) because it does not rely on correct specification of the conditional density of $X_i$, given $Z_i$.

## 2.3 Exact estimation in probit structural mean models

The previous estimation principles can also be used for fitting the probit structural mean model

$$\Phi^{-1}\left\{\mathrm{E}(Y_i|X_i,Z_i)\right\}-\Phi^{-1}\left\{\mathrm{E}(Y_{i0}|X_i,Z_i)\right\}=\phi^*X_i, \tag{3.12}$$

where $\Phi^{-1}$ is the probit link and $\phi^*$ is unknown. This is possible upon replacing the logit link with the probit link. A simpler parametric estimator can be obtained under the additional assumptions:

1. that the exposure is normally distributed conditional on the instrumental variable with mean $\alpha_0^*+\alpha_1^*Z_i$ and constant standard deviation $\sigma^*$, where $\alpha_0^*,\alpha_1^*,\sigma^*$ are unknown;

2. that a probit association model holds:

$$\Phi^{-1}\left\{\mathrm{E}(Y_i|X_i,Z_i)\right\}=\theta_0^*+\theta_1^*X_i+\theta_2^*Z_i, \tag{3.13}$$

   where $\theta_0^*,\theta_1^*,\theta_2^*$ are unknown.

Indeed, combining the probit structural mean model (3.12) and association model (3.13) yields

$$\mathrm{E}(Y_{i0}|X_i,Z_i)=\Phi\{\theta_0^*+(\theta_1^*-\phi^*)X_i+\theta_2^*Z_i\}.$$

Averaging over the exposure, conditional on $Z_i$ (see the Appendix), then gives

$$\mathrm{E}(Y_{i0}|Z_i)=\Phi\left\{\frac{\theta_0^*+\theta_2^*Z_i+(\theta_1^*-\phi^*)(\alpha_0^*+\alpha_1^*Z_i)}{\sqrt{1+(\theta_1^*-\phi^*)^2\sigma^{2*}}}\right\}, \tag{3.14}$$

from which it follows that $\theta_2^*=(\phi^*-\theta_1^*)\alpha_1^*$. This suggests the following two-stage approach:

1. First regress the exposure on the instrumental variable using ordinary least squares and obtain estimates $\hat{\alpha}_1$ for the regression slope $\alpha_1^*$ and $\hat{\sigma}$ for the residual standard deviation $\sigma^*$;

2. Next regress outcome on exposure and instrumental variable by fitting model (3.13) to obtain estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ for the regression slopes $\theta_1^*$ and $\theta_2^*$, respectively.

The resulting estimator

$$\hat{\phi}=\hat{\theta}_1+\frac{\hat{\theta}_2}{\hat{\alpha}_1} \tag{3.15}$$

for $\phi^*$ will be referred to as 'Two-stage IV-estimator I' throughout.

An alternative estimator is obtained by averaging over the exposure in the association model (3.13) and using the previous identity $\theta_2^* = (\phi^* - \theta_1^*)\alpha_1^*$ to obtain

$$\mathrm{E}(Y_i|Z_i) = \Phi\left(\frac{\theta_0^* + \phi^*\alpha_1^* Z_i + \theta_1^*\alpha_0^*}{\sqrt{1 + \theta_1^{*2}\sigma^{2*}}}\right).$$

This suggests regressing the outcome on the instrumental variable using the probit regression model

$$\Phi^{-1}\{\mathrm{E}(Y_i|Z_i)\} = \lambda_0^* + \lambda_1^* Z_i \tag{3.16}$$

to obtain an estimate $\hat{\lambda}_1$ for the unknown regression slope $\lambda_1^*$, and then estimating $\phi^*$ as

$$\hat{\phi} = \frac{\hat{\lambda}_1\sqrt{1 + \hat{\theta}_1^2\hat{\sigma}^2}}{\hat{\alpha}_1}. \tag{3.17}$$

We will refer this estimator as the 'Two-stage IV-estimator II' throughout. Both these estimators have, to the best of our knowledge, not been previously considered.

Thomas and Conti (2004) propose to estimate $\phi^*$ as

$$\hat{\phi} = \frac{\hat{\lambda}_1}{\sqrt{\hat{\alpha}_1^2 - \hat{\lambda}_1^2\hat{\sigma}^2}}.$$

This estimator will be referred to as 'Two-stage IV-estimator III' throughout. It is obtained by substituting $\hat{\theta}_1$ with $\hat{\phi}$ in expression (3.17).

An advantage of Two-stage IV-estimators II and III over Two-stage IV-estimator I is that they preserve the Type I error rate of tests of the causal null hypothesis (i.e. $\phi^* = 0$) when the model is misspecified, because $\lambda_1^* = 0$ at the causal null hypothesis. An additional advantage of Two-stage IV-estimator III is that it does not involve estimates from the two association models (3.13) and (3.16) and may therefore lend itself better to use in meta-analyses based on summary statistics. However, we do not recommend Two-stage IV-estimator III for data analysis because it is not guaranteed to be be valid when the exposure effect on outcome differs from zero (i.e. $\phi^* = 0$) and is confounded (i.e. $\theta_2^* \neq 0$). In addition, this estimator is only defined when $-\hat{\alpha}_1/\hat{\sigma} < \hat{\lambda}_1 < \hat{\alpha}_1/\hat{\sigma}$.

When the outcome mean lies between 10% and 90%, the above Two-stage IV-estimators yield approximate estimates of the causal odds ratio because of the identity $\exp(\psi^*) \approx \exp(\phi^*/0.6071)$ (McCullagh and Nelder, 1983). For dichotomous exposures, related estimators can be obtained via probit structural equation models

that replace the linear regression model for $X_i$ in assumption 1 above, with a probit regression model (see e.g. Rassen et al., 2008).

## 2.4 Approximate estimation in logistic structural mean models

Approximate IV-estimators of the causal odds ratio can be obtained by averaging over the observed exposure values in model (3.2) using the following approximations

$$\begin{aligned} \mathrm{E}\left\{\text{logit } \mathrm{E}(Y_i|X_i, Z_i)|Z_i\right\} &\approx \text{logit } \mathrm{E}(Y_i|Z_i) \\ \mathrm{E}\left\{\text{logit } \mathrm{E}(Y_{i0}|X_i, Z_i)|Z_i\right\} &\approx \text{logit } \mathrm{E}(Y_{i0}|Z_i). \end{aligned}$$

This together with the logistic structural mean model (3.2) implies

$$\begin{aligned} \text{logit } \mathrm{E}(Y_i|Z_i) &\approx \text{logit } \mathrm{E}(Y_{i0}|Z_i) + \psi^* E(X_i|Z_i) \\ &= \omega^* + \psi^* E(X_i|Z_i), \end{aligned} \tag{3.18}$$

where $\omega^* \equiv \text{logit}\mathrm{E}(Y_{i0}) = \text{logit}\mathrm{E}(Y_{i0}|Z_i)$ under the IV-assumptions. An approximate IV-estimator of the causal odds ratio may thus be obtained using the following two-stage approach:

1. First obtain an estimate of the expected exposure in function of the IV by fitting an appropriate regression model. Let the predicted exposure be $\hat{X}_i \equiv \hat{\mathrm{E}}(X_i|Z_i)$.

2. Next, fit the logistic regression model

$$\text{logit } \mathrm{E}(Y_i|Z_i) = \omega + \psi \hat{X}_i. \tag{3.19}$$

   The estimate of the regression slope in that model will be referred to as the 'Approximate IV-estimator' of $\psi^*$ throughout.

When the IV is dichotomous, it follows from (3.18) that

$$OR_{Y|Z} \equiv \frac{\text{odds}(Y_i = 1|Z_i = 1)}{\text{odds}(Y_i = 1|Z_i = 0)} \approx \exp(\psi^*)^{\Delta_{X|Z}}$$

where $\Delta_{X|Z} \equiv \mathrm{E}(X_i|Z_i = 1) - \mathrm{E}(X_i|Z_i = 0)$, or equivalently,

$$\psi^* \approx \frac{\log(OR_{Y|Z})}{\Delta_{X|Z}}. \tag{3.20}$$

The Approximate IV-estimator is commonly employed in the analysis of Mendelian randomization studies (Thompson et al., 2003) and comes under a variety of names (e.g. a Wald-type estimator in Didelez, Meng and Sheehan (2008) and 2-stage logistic approach in Rassen et al. (2008)). It lends itself particularly well to use in meta-analyses based on summary statistics because the approximation (3.20) can be used, even when information on $OR_{Y|Z}$ and $\Delta_{X|Z}$ is obtained from different studies (Minelli et al., 2004; Smith et al., 2005). However, because of the made approximations, the Approximate IV-estimator tends to be biased, even in large samples, and requires correct specification of the first stage regression model for the expected exposure (Didelez, Meng and Sheehan, 2008; Rassen et al., 2008; Henneman, van der Laan and Hubbard, 2002). This is true except at the null hypothesis of no causal effect because $Y_i \perp\!\!\!\perp Z_i$ at the null hypothesis so that the usual maximum likelihood estimator of $\psi$ indexing model (3.19) will then converge to 0 in probability.

The bias of the Approximate IV-estimator can sometimes be attenuated by additionally including $R_i \equiv X_i - \hat{X}_i$ as a regressor in model (3.19) (Nagelkerke et al., 2000; Palmer et al., 2008). The resulting estimator of $\psi^*$ will be referred to as the 'Adjusted IV-estimator'. This modification often tends to be rewarding because $R_i$ captures part of the confounders that influence the relationship between $X_i$ and $Y_i$, so that adjustment for $R_i$ removes some residual confounding bias. Indeed, suppose that all confounders of the exposure effect can be summarized in a scalar measurement $U_i$ and that the contributions of the instrument $Z_i$ and confounder $U_i$ are additive in the sense that $X_i = h(Z_i) + U_i$ for some function $h$. Suppose additionally that the conditional mean $E(X_i|Z_i)$ is known so that $\hat{X}_i = h(Z_i)$ and thus $R_i = U_i$. Then fitting model

$$\begin{aligned} \text{logit E}(Y_i|X_i, Z_i, U_i) &= \tilde{\beta}_0^* + \psi^* X_i + \tilde{\beta}_1^* R_i \\ &= \tilde{\beta}_0^* + \psi^* E(X_i|Z_i) + (\psi^* + \tilde{\beta}_1^*) R_i \end{aligned}$$

will yield a consistent estimator of the conditional causal odds ratio

$$\psi^* = \frac{\text{odds}\,(Y_i = 1|X_i = 1, Z_i, U_i)}{\text{odds}\,(Y_{i0} = 1|X_i = 1, Z_i, U_i)} = \frac{\text{odds}\,(Y_i = 1|X_i = 1, Z_i)}{\text{odds}\,(Y_{i0} = 1|X_i = 1, Z_i)}, \tag{3.21}$$

where we use that $U_i$ is completely determined by $X_i$ and $Z_i$ in the last identity. When the contributions of the instrument $Z_i$ and confounder $U_i$ on the exposure are not additive, then the Approximate IV-estimator may be biased, even at the causal null hypothesis. In contrast, when they are additive, then it is still prone to some bias because the model for $E(X_i|Z_i)$ may be misspecified and because, even when it is correctly specified, $E(X_i|Z_i)$ is not known in practice so that $R_i$ is an imprecise estimate of $U_i$. It remains to be explored whether methods for measurement

error correction, such as SIMulation-EXtrapolation (Carroll et al., 2006), can help attenuate this bias.

## 2.5 Generalized method of moments

In the econometrics literature, the confounder $U_i$ is commonly assumed to have an additive effect on the outcome (Amemiya, 1974; Foster, 1997; Johnston et al., 2008) in the sense that

$$\mathrm{E}(Y_i|X_i, U_i) = \mathrm{expit}(\beta^* + \tilde{\psi}^* X_i) + U_i, \tag{3.22}$$

where $\beta^*, \tilde{\psi}^*$ are unknown and where $\mathrm{E}(U_i|X_i) = 0$. This model implies that

$$\mathrm{logit}\mathrm{E}(Y_i|X_i, U_i) - \mathrm{logit}\mathrm{E}(Y_{i0}|X_i, U_i) = \tilde{\psi}^* X_i, \tag{3.23}$$

which is closely related to the logistic structural mean model (3.2), except that it additionally conditions on $U_i$. Because, for each $x$, $Y_{ix} \perp\!\!\!\perp X_i|U_i$ when $U_i$ represents all confounders of the exposure effect, model (3.23) implies the marginal structural model

$$\mathrm{E}(Y_{ix}) = \mathrm{E}\left\{\mathrm{E}(Y_{ix}|X_i = x, U_i)\right\} = \mathrm{expit}(\beta^* + \tilde{\psi}^* x)$$

considered by Henneman, van der Laan and Hubbard (2002). This demonstrates that $\exp(\tilde{\psi}^*)$ can be interpreted as the marginal (i.e. population averaged) causal odds ratio

$$\exp(\tilde{\psi}^*) = \frac{\mathrm{odds}\left(Y_{i1} = 1\right)}{\mathrm{odds}\left(Y_{i0} = 1\right)}.$$

This may differ from (3.1) because of noncollapsibility of the odds ratio and because subjects with different observed exposure levels may experience different effects of the same exposure. Using that $Z_i \perp\!\!\!\perp U_i$ under the IV-assumptions, estimators $\hat{\beta}$ for $\beta^*$ and $\hat{\psi}$ for $\psi^*$ can be obtained by solving the following unbiased estimating equations:

$$\begin{aligned} 0 &= \sum_{i=1}^{n} Y_i - \mathrm{expit}(\beta + \psi X_i) \\ 0 &= \sum_{i=1}^{n} Z_i \left\{Y_i - \mathrm{expit}(\beta + \psi X_i)\right\}. \end{aligned}$$

An efficient estimator is obtained (Greene, 2003) by next calculating the $2 \times 2$ matrix

$$W \equiv \frac{1}{n} \sum_{i=1}^{n} (1 \ Z_i)' \, e_i^2 \, (1 \ Z_i)$$

with $e_i \equiv Y_i - \text{expit}(\hat{\beta} + \hat{\psi} X_i)$ and subsequently re-estimating $\beta^*$ and $\psi^*$ as those values minimizing

$$\sum_{i=1}^{n} (1 \; Z_i) \, e_i^2 W^{-1} \, (1 \; Z_i)' \, . \tag{3.24}$$

The resulting estimator will be referred to as the 'GMM estimator' throughout.

It follows from the unbiasedness of the estimating functions $(1 \; Z_i)' \{Y_i - \text{expit}(\beta^* + \tilde{\psi}^* X_i)\}$ at $\tilde{\psi}^* = 0$ that the GMM estimator is (locally) robust against model misspecification at the null hypothesis of no causal effect. However, it is not guaranteed to exist and is inconsistent away from the causal null hypothesis because the dichotomous nature of the outcome implies that the error term $U_i$ retains a dependence on $X_i$ so that the basic assumption $Z_i \perp\!\!\!\perp U_i$ underlying this method is violated (Henneman, van der Laan and Hubbard, 2002).

# 3 Application

Studies of outcomes associated with exposure to pharmaceutical products in routine clinical practice are often observational. In this section, we analyse one such study (Brookhart et al., 2006; Brookhart and Schneeweiss, 2007) where the goal is to assess short-term effects of Cox-2 treatment (as compared to non-steroidal anti-inflammatory treatment) on the risk of gastrointestinal (GI) bleeding within 60 days. As Table 3.1 shows, of the 37 842 new non-selective NSAID users drawn from a large population based cohort of medicare beneficiaries who were eligible for a state-run pharmaceutical benefit plan, 26 407 patients were placed on Cox-2 treatment. Let the received treatment $X_i$ equal 1 if subject $i$ was placed on Cox-2 and 0 for non-selective NSAIDs. Let the outcome $Y_i$ indicate 1 for upper gastrointestinal (GI) bleeding within 60 days of initiating an NSAID for subject $i$, and 0 otherwise. As in Brookhart and Schneeweiss (2007), we use the physician's prescribing preference for Cox-2 (versus non-selective NSAIDs) $Z_i$ as an instrumental variable for the effect of Cox-2 treatment on gastrointestinal bleeding. To obtain Exact IV-estimator I under model (3.2), we first fitted the logistic association model

$$\text{logit}\hat{\text{E}}\,(Y_i | X_i, Z_i) = -4.89 + 0.11 X_i - 0.33 Z_i$$

and then obtained $\hat{\psi} = -2.508$ by solving (3.5). This corresponds with a causal odds ratio of $\exp(\hat{\psi}) = 0.081$ (95% confidence interval $(0.010, 0.826)$). The same result is obtained using Exact IV-estimator II. This result is in stark contrast with the logistic regression estimate $\exp(0.11) = 1.12$ (95% confidence interval $(0.849, 1.495)$)

Table 3.1: *Observed data with $X_i$ indicating received treatment (Cox-2 (1) versus non-selective NSAIDs (0)), $Z_i$ indicating the physician's prescribing preference (Cox-2 (1) versus non-selective NSAIDs (0)), and $Y_i$ indicating gastrointestinal (GI) bleeding (1) within 60 days of initiating an NSAID for subject $i$.*

| | $Z_i = 0$ | | $Z_i = 1$ | |
|---|---|---|---|---|
| | $Y_i = 0$ | $Y_i = 1$ | $Y_i = 0$ | $Y_i = 1$ |
| $X_i = 0$ | 5640 | 39 | 5722 | 34 |
| $X_i = 1$ | 6740 | 60 | 19493 | 114 |

as obtained from the above association model. To obtain the Two-stage IV-estimators, we next fitted a probit model of $Y_i$ on $X_i$ and $Z_i$

$$\Phi^{-1}\left\{\mathrm{E}(Y_i|X_i, Z_i)\right\} = -2.43 + 0.04X_i - 0.12Z_i.$$

Further noting that a linear regression analysis of $X_i$ on $Z_i$ yields an estimated slope of $\hat{\alpha_1} = 0.23$ and a residual variance of $\hat{\sigma}^2 = 0.20$, we obtain a causal odds ratio of 0.454 (95% confidence interval $(0.240, 0.870)$) using Two-stage IV-estimator I, 0.452 (95% confidence interval $(0.234, 0.908)$) using Two-stage IV-estimator II and 0.443 (95% confidence interval $(0.197, 0.924)$) using Two-stage IV-estimator III. Given the small risk of gastrointestinal bleeding (0.65%), these Two-stage IV-estimates are likely biased estimates of the causal odds ratio. Finally, from a logistic regression of $Y_i$ on $Z_i$,

$$\mathrm{logit}\mathrm{E}(Y_i|Z_i) = -4.83 - 0.31Z_i,$$

we obtain a causal odds ratio of $\exp(-0.31/0.23) = 0.258$ (95% confidence interval $(0.086, 0.798)$) using the Approximate IV-estimator. By additionally adjusting for the residual in a linear regression analysis of $X_i$ on $Z_i$, a very similar estimate of 0.251 (95% confidence interval $(0.085, 0.758)$) is obtained using the Adjusted IV-estimator. Finally, the GMM-estimator (0.401 with 95% confidence interval $(0.083, 2.422)$) should not be well trusted because the objective function (3.24) reaches a minimum far from zero. The results of the data analysis are summarized in Table 3.2. Overall, on the basis of Exact IV-estimator I, we estimate roughly that for new NSAID users on Cox-2, the risk of gastrointestinal bleeding would increase with at least 17% $(= 1 - 0.826)$ if they were to use non-selective NSAIDs. Note that because of the low prevalence of gastrointestinal bleeding, the different estimates have large imprecision, despite the large sample size. To improve our understanding of the relative performance of the different estimators, we will conduct simulation experiments in the next section.

Table 3.2: *Estimated causal odds ratios and 95% bootstrap percentile confidence intervals based on 10 000 bootstrap resamples.*

| IV-Estimators | $\exp(\hat{\psi})$ | 95%CI |
|---|---|---|
| Logist. reg. | 1.121 | (0.849, 1.495) |
| Exact I | 0.081 | (0.010, 0.826) |
| Approximate | 0.258 | (0.086, 0.798) |
| Adjusted | 0.251 | (0.085, 0.758) |
| Two-stage I | 0.453 | (0.240, 0.870) |
| Two-stage II | 0.452 | (0.234, 0.908) |
| Two-stage III | 0.443 | (0.197, 0.924) |

# 4 Simulation study

## 4.1 Dichotomous exposure

To compare the performance of the considered IV-estimators in the presence of unmeasured confounders, we conducted several simulation studies. Each simulation study was based on 1000 replications of sample size 1000. In each replication, a dichotomous instrumental variable $Z_i$ was generated with $P(Z_i = 0) = P(Z_i = 1) = 0.50$. Next, a dichotomous exposure $X_i$ was generated, which takes the value 1 with probability $P(X_i = 1|Z_i) = \text{expit}(\alpha_0^* + \alpha_1^* Z_i)$, for chosen values of $\alpha_0^*$ and $\alpha_1^*$, and 0 otherwise. Finally, a dichotomous outcome was generated under the logistic structural mean model (3.2) with the additional assumption that $q(X_i, Z_i; \eta^*) = \eta^* X_i$, for chosen values of $\psi^*$ and $\eta^*$ (additional simulations with $q(X_i, Z_i; \eta^*) = \eta_0^* X_i + \eta_1^* X_i Z_i$ gave qualitatively similar results). This happened by letting $Y_i$ equal 1 with probability

$$P(Y_i = 1|X_i, Z_i) = \text{expit}\{(\psi^* + \eta^*)X_i + v(Z_i; \eta^*, \omega^*)\}$$

and 0 otherwise, where $v(Z_i; \eta^*, \omega^*)$ solves

$$\text{E}(Y_{i0}) = \sum_{x=0}^{1} \text{expit}\{\eta^* x + v(Z_i; \eta^*, \omega^*)\} P(X_i = x|Z_i),$$

and where $\text{E}(Y_{i0})$ was chosen to match settings with either high or low prevalence $E(Y_i)$. In particular, we chose the mean outcome $\text{E}(Y_i)$ to equal $0.05, 0.25$ or $0.50$, causal effect sizes $\psi^* = 0$ or 2, and different degrees of confounding $\eta^*$ equal to $-2$

or 0, where $\eta^* = 0$ encodes no confounding of the effect of $X$ on $Y$. These different parameter settings represent studies with instrumental variables of varying strength, corresponding to mean differences $\Delta_{X|Z}$ equal to 0.15 or 0.25, and corresponding correlation coefficients between $X$ and $Z$ equal to 0.16 and 0.26, respectively. In the case where the information was weakest (mean outcome 5% and $\delta_{xy} = 0.15$) we additionally considered a sample size of 10 000.

Amongst the exact estimation approaches, the closed-form estimator (3.5) was evaluated under the nonparametric working association model $m(X_i, Z_i; \beta^*) = \beta_0^* + \beta_1^* X_i + \beta_2^* Z_i + \beta_3^* X_i Z_i$. When the discriminant $Q_1^2 - 4Q_2Q_3(Q_2 - \hat{X}_{11} + \hat{X}_{10})$ in (3.6) was negative, no estimator was obtained with this approach. In the rare occasions where 2 solutions were obtained, the one closest to the Two-stage IV-estimator I was selected because of its relatively adequate performance as compared to the other approximate estimators. Exact IV-estimator II was not displayed because it was identical to Exact IV-estimator I. Finally, no Two-stage IV-estimator III was obtained when $\hat{\alpha}_1^2 - \hat{\lambda}_1^2 \hat{\sigma}^2$ was negative.

Tables 3.3-3.5 present the results of the simulation studies for dichotomous exposure, including the ordinary logistic regression estimator. Because of outlying estimates in a number of simulations, we focus on the median bias (mean bias), the robust MCD-estimate of the standard deviation (sample standard deviation), the median absolute deviation (i.e. median$|\hat{\psi} - \psi^*|$) (mean squared error), and finally the percentage of simulations in which an estimate was obtained (% conv.).

The simulation analyses in Table 3.3 reveals that only the Exact IV-estimator is guaranteed to be asymptotically unbiased, although there can be a bias when the outcome prevalence is small. This bias was no longer seen in a simulation study with sample size 10 000 (not shown), indicating that it is a finite-sample bias due to the lack of information at low prevalence. All other estimators suffer bias, even when the association between outcome and exposure is unconfounded (i.e. $\eta^* = 0$), but tend to be less vulnerable to outlying estimates. Disregarding outliers (i.e. considering the robust MCD estimator), all estimators have a variability of similar magnitude (except for Two-stage IV-estimator III and the GMM-estimator, which tend to vary more).

Results in Table 3.4 confirm, in line with the theory, that the Exact IV-estimator, Two-stage IV-estimators II and III, the Approximate IV-estimator and GMM-estimator are unbiased in the absence of a causal effect, but that the other estimators can have substantial bias at the causal null hypothesis. In line with Palmer et al. (2008), we observe a larger bias of the Approximate IV-estimator as the association between $Y$ and $X$ becomes more positive, and thus as $\eta^*$ increases. We observed no clear benefit of the Adjusted IV-estimator (and often even a somewhat worse performance). The Two-stage IV-estimator I was closest to the Adjusted IV-estimator, often yielding cor-

relations above 0.99, and was generally doing better than the Two-stage IV-estimator II and III, except at the causal null hypothesis.

Table 3.3: *Simulation results when exposure $X$ is dichotomous, $\psi^* = 2$ and $\delta_{xz} = 0.25$.*

| | | | Bias | SE | | |
|---|---|---|---|---|---|---|
| E(**Y**) | ($\eta^*$, $\omega^*$) | **Methods** | **Median (Mean)** | **MCD (Sample)** | **MAD (MSE)** | **% conv.** |
| 0.05 | (-2, -3.90) | Exact I | -0.54 (-0.29) | 0.79 (1.76) | 0.82 (3.20) | 71.0 |
| | | Approximate | 0.21 (0.25) | 1.11 (1.25) | 0.81 (1.62) | 100 |
| | | Adjusted | 0.22 (0.26) | 1.11 (1.25) | 0.82 (1.63) | 100 |
| | | Logist. reg. | -1.81 (-1.81) | 0.29 (0.33) | 1.81 (3.40) | 100 |
| | | Two-stage I | -0.31 (-0.29) | 0.83 (0.93) | 0.67 (0.95) | 100 |
| | | Two-stage II | -0.32 (-0.30) | 0.83 (0.93) | 0.68 (0.95) | 100 |
| | | Two-stage III | -0.15 (0.37) | 1.02 (2.59) | 0.83 (6.86) | 98.0 |
| | | GMM | -0.65 (-0.44) | 1.05 (1.92) | 1.07 (3.90) | 29.4 |
| 0.05 | (0, -4.55) | Exact I | -0.45 (0.02) | 1.10 (3.67) | 0.92 (13.47) | 79.1 |
| | | Approximate | -0.75 (-0.74) | 1.00 (1.15) | 0.91 (1.87) | 100 |
| | | Adjusted | -0.45 (-0.20) | 1.06 (2.11) | 0.87 (4.50) | 100 |
| | | Logist. reg. | 0.04 (0.49) | 0.51 (2.54) | 0.40 (6.70) | 100 |
| | | Two-stage I | -0.54 (-0.40) | 0.84 (1.31) | 0.75 (1.88) | 100 |
| | | Two-stage II | -0.97 (-0.94) | 0.83 (1.00) | 1.03 (1.89) | 100 |
| | | Two-stage III | -1.03 (-0.91) | 0.83 (1.14) | 1.10 (2.13) | 99.4 |
| | | GMM | -1.20 (-1.01) | 0.98 (1.26) | 1.34 (2.63) | 61.9 |
| 0.25 | (-2, -2.13) | Exact I | -0.03 (0.11) | 0.54 (0.79) | 0.43 (0.63) | 95.1 |
| | | Approximate | 0.35 (0.39) | 0.61 (0.67) | 0.50 (0.60) | 100 |
| | | Adjusted | 0.35 (0.39) | 0.61 (0.67) | 0.50 (0.60) | 100 |
| | | Logist. reg. | -1.82 (-1.82) | 0.15 (0.16) | 1.82 (3.34) | 100 |
| | | Two-stage I | 0.28 (0.32) | 0.59 (0.65) | 0.45 (0.52) | 100 |
| | | Two-stage II | 0.27 (0.32) | 0.59 (0.65) | 0.45 (0.52) | 100 |
| | | Two-stage III | 0.87 (1.57) | 0.98 (3.57) | 0.94 (15.22) | 97.6 |
| | | GMM | 0.62 (0.81) | 0.97 (1.65) | 0.80 (3.38) | 28.3 |
| 0.25 | (0 , -2.65) | Exact I | -0.05 (0.02) | 0.68 (0.91) | 0.51 (0.83) | 96.5 |
| | | Approximate | -0.52 (-0.54) | 0.51 (0.57) | 0.57 (0.62) | 100 |
| | | Adjusted | -0.53 (-0.54) | 0.57 (0.63) | 0.60 (0.70) | 100 |
| | | Logist. reg. | 0.01 (0.02) | 0.21 (0.24) | 0.16 (0.06) | 100 |
| | | Two-stage I | -0.20 (-0.19) | 0.56 (0.62) | 0.43 (0.42) | 100 |
| | | Two-stage II | -0.39 (-0.41) | 0.56 (0.62) | 0.52 (0.56) | 100 |
| | | Two-stage III | -0.44(-0.37) | 0.61 (0.85) | 0.63 (0.86) | 100 |
| | | GMM | -0.29 (-0.08) | 0.80 (1.11) | 0.71 (1.24) | 84.4 |
| 0.50 | (-2, -1.15) | Exact I | -0.01 (0.03) | 0.37 (0.47) | 0.27 (0.22) | 100 |
| | | Approximate | 0.45 (0.50) | 0.53 (0.61) | 0.50 (0.62) | 100 |
| | | Adjusted | 0.45 (0.51) | 0.54 (0.61) | 0.51 (0.63) | 100 |
| | | Logist. reg. | -1.82 (-1.82) | 0.13 (0.13) | 1.82 (3.31) | 100 |
| | | Two-stage I | 0.53 (0.58) | 0.55 (0.63) | 0.56 (0.73) | 100 |
| | | Two-stage II | 0.53 (0.58) | 0.55 (0.63) | 0.56 (0.73) | 100 |
| | | Two-stage III | 1.42 (2.34) | 1.05 (4.77) | 1.42 (28.27) | 94.8 |
| | | GMM | 0.86 (1.07) | 0.78 (1.27) | 0.87 (2.77) | 80.5 |
| 0.50 | (0, -1.40) | Exact I | 0.04 (0.02) | 0.52 (0.60) | 0.39 (0.36) | 100 |
| | | Approximate | -0.15 (-0.18) | 0.41 (0.47) | 0.32 (0.25) | 100 |
| | | Adjusted | -0.45 (-0.48) | 0.50 (0.56) | 0.50 (0.54) | 100 |
| | | Logist. reg. | -0.005 (0.005) | 0.15 (0.16) | 0.11 (0.03) | 100 |
| | | Two-stage I | 0.06 (0.03) | 0.50 (0.56) | 0.37 (0.31) | 100 |
| | | Two-stage II | 0.18 (0.14) | 0.49 (0.55) | 0.40 (0.32) | 100 |
| | | Two-stage III | 0.23 (0.30) | 0.69 (0.88) | 0.52 (0.86) | 99.9 |
| | | GMM | 0.05 (0.10) | 0.59 (0.75) | 0.45 (0.57) | 99.2 |

## 4.2 Normally distributed exposure

We conducted a second simulation experiment whereby the exposure $X_i$ was drawn from a normal distribution with mean $\alpha_0^* + \alpha_1^* Z_i$ and constant variance 1. We made the same choices of the outcome mean $\mathrm{E}(Y_i)$, selected $\psi^*$ equal to 0 or 1 and $\eta^*$

Table 3.4: *Simulation results when exposure $X$ is dichotomous, $\psi^* = 0$ and $\delta_{xz} = 0.25$.*

| | | | **Bias** | **SE** | | |
|---|---|---|---|---|---|---|
| E(**Y**) | ($\eta^*$, $\omega^*$) | **Methods** | **Median (Mean)** | **MCD (Sample)** | **MAD (MSE)** | % **conv.** |
| 0.05 | (-2, -2.90) | Exact I | -0.23 (-0.28) | 0.88 (1.36) | 0.71 (1.94) | 83.7 |
| | | Approximate | -0.03 (0.06) | 1.05 (1.20) | 0.77 (1.44) | 100 |
| | | Adjusted | 0.63 (0.54) | 1.09 (1.26) | 0.91 (1.99) | 100 |
| | | Logist. reg. | -1.84 (-1.85) | 0.30 (0.33) | 1.84 (3.55) | 100 |
| | | Two-stage I | 0.22 (0.30) | 0.90 (1.00) | 0.68 (1.10) | 100 |
| | | Two-stage II | -0.02 (0.05) | 0.87 (0.99) | 0.63 (0.99) | 100 |
| | | Two-stage III | -0.02 (0.06) | 0.80 (1.06) | 0.59 (1.12) | 100 |
| | | GMM | -0.15 (-0.01) | 0.90 (1.42) | 0.67 (2.01) | 89.5 |
| 0.05 | (0, -2.90) | Exact I | -0.02 (-0.17) | 1.00 (1.40) | 0.73 (1.99) | 95.0 |
| | | Approximate | 0.02 (0.005) | 1.03 (1.14) | 0.76 (1.30) | 100 |
| | | Adjusted | 0.02 (0.01) | 1.03 (1.14) | 0.76 (1.30) | 100 |
| | | Logist. reg. | 0.02 (0.02) | 0.29 (0.32) | 0.22 (0.10) | 100 |
| | | Two-stage I | 0.02 (0.02) | 0.79 (0.87) | 0.58 (0.75) | 100 |
| | | Two-stage II | 0.02 (0.004) | 0.79 (0.86) | 0.58 (0.75) | 100 |
| | | Two-stage III | 0.02 (0.003) | 0.79 (0.97) | 0.59 (0.94) | 100 |
| | | GMM | -0.06 (0.007) | 0.96 (1.26) | 0.71 (1.60) | 92.4 |
| 0.25 | (-2, -1.10) | Exact I | -0.02 (0.001) | 0.40 (0.48) | 0.30 (0.23) | 100 |
| | | Approximate | -0.03 (0.01) | 0.54 (0.60) | 0.40 (0.36) | 100 |
| | | Adjusted | 0.87 (0.93) | 0.65 (0.74) | 0.87 (1.40) | 100 |
| | | Logist. reg. | -1.80 (-1.80) | 0.14 (0.16) | 1.80 (3.26) | 100 |
| | | Two-stage I | 0.38 (0.42) | 0.60 (0.69) | 0.52 (0.65) | 100 |
| | | Two-stage II | -0.03 (0.01) | 0.59 (0.66) | 0.44 (0.44) | 100 |
| | | Two-stage III | -0.03 (0.01) | 0.53 (0.61) | 0.39 (0.37) | 100 |
| | | GMM | -0.03 (0.06) | 0.52 (0.69) | 0.40 (0.49) | 97.7 |
| 0.25 | (0, -1.10) | Exact I | 0.006 (-0.02) | 0.52 (0.59) | 0.40 (0.35) | 100 |
| | | Approximate | 0.006 (-0.01) | 0.54 (0.60) | 0.40 (0.36) | 100 |
| | | Adjusted | 0.01 (-0.01) | 0.53 (0.60) | 0.40 (0.36) | 100 |
| | | Logist. reg. | 0.002 (0.0003) | 0.14 (0.16) | 0.11 (0.02) | 100 |
| | | Two-stage I | 0.007 (-0.01) | 0.52 (0.58) | 0.38 (0.34) | 100 |
| | | Two-stage II | 0.006 (-0.01) | 0.52 (0.58) | 0.38 (0.34) | 100 |
| | | Two-stage III | 0.006 (-0.01) | 0.52 (0.61) | 0.39 (0.37) | 100 |
| | | GMM | 0.006 (0.03) | 0.56 (0.65) | 0.40 (0.42) | 100 |
| 0.50 | (-2, 0) | Exact I | -0.03 (-0.03) | 0.36 (0.40) | 0.26 (0.16) | 100 |
| | | Approximate | -0.04 (0.01) | 0.44 (0.51) | 0.33 (0.26) | 100 |
| | | Adjusted | 0.62 (0.67) | 0.54 (0.62) | 0.63 (0.83) | 100 |
| | | Logist. reg. | -1.82 (-1.81) | 0.14 (0.16) | 1.82 (3.32) | 100 |
| | | Two-stage I | 0.10 (0.14) | 0.53 (0.61) | 0.40 (0.39) | 100 |
| | | Two-stage II | -0.04 (0.01) | 0.52 (0.60) | 0.39 (0.36) | 100 |
| | | Two-stage III | -0.04 (0.01) | 0.46 (0.55) | 0.35 (0.30) | 100 |
| | | GMM | -0.04 (0.01) | 0.44 (0.53) | 0.33 (0.28) | 100 |
| 0.50 | (0, 0) | Exact I | -0.03 (-0.03) | 0.48 (0.50) | 0.36 (0.25) | 100 |
| | | Approximate | -0.03 (-0.03) | 0.49 (0.52) | 0.36 (0.27) | 100 |
| | | Adjusted | -0.01 (-0.03) | 0.49 (0.52) | 0.37 (0.27) | 100 |
| | | Logist. reg. | -0.004 (-0.004) | 0.12 (0.14) | 0.09 (0.02) | 100 |
| | | Two-stage I | -0.03 (-0.03) | 0.50 (0.54) | 0.37 (0.29) | 100 |
| | | Two-stage II | -0.03 (-0.03) | 0.50 (0.54) | 0.37 (0.29) | 100 |
| | | Two-stage III | -0.03 (-0.03) | 0.51 (0.56) | 0.37 (0.31) | 100 |
| | | GMM | -0.03 (-0.03) | 0.49 (0.54) | 0.36 (0.29) | 100 |

Table 3.5: *Simulation results when exposure X is dichotomous,* $\psi^* = 2$ *and* $\delta_{xz} = 0.15$.

| | | | **Bias** | **SE** | | |
|---|---|---|---|---|---|---|
| E(**Y**) | ($\eta^*$, $\omega^*$) | **Methods** | **Median (Mean)** | **MCD (Sample)** | **MAD (MSE)** | % **conv.** |
| 0.05 | (-2, -3.75) | Exact I | -1.08 (-1.19) | 1.15 (1.62) | 1.26 (4.03) | 62.8 |
| | | Approximate | 0.01 (0.10) | 1.87 (2.20) | 1.36 (4.85) | 100 |
| | | Adjusted | 0.01 (0.11) | 1.86 (2.20) | 1.36 (4.84) | 100 |
| | | Logist. reg. | -1.95 (-1.94) | 0.28 (0.31) | 1.95 (3.86) | 100 |
| | | Two-stage I | -0.48 (-0.40) | 1.40 (1.66) | 1.06 (2.91) | 100 |
| | | Two-stage II | -0.46 (-0.41) | 1.41 (1.66) | 1.05 (2.91) | 100 |
| | | Two-stage III | -0.54 (0.02) | 1.52 (3.20) | 1.39 (10.23) | 88.6 |
| | | GMM | -1.48 (-1.38) | 1.30 (1.83) | 1.58 (5.25) | 42.1 |
| 0.05 | (0, -4.50) | Exact I | -0.93 (-0.81) | 1.59 (3.70) | 1.37 (14.37) | 70.6 |
| | | Approximate | -0.73 (-0.65) | 1.71 (2.05) | 1.36 (4.64) | 100 |
| | | Adjusted | -0.57 (-0.35) | 1.75 (2.46) | 1.34 (6.19) | 100 |
| | | Logist. reg. | 0.02 (0.29) | 0.46 (1.89) | 0.34 (3.66) | 100 |
| | | Two-stage I | -0.56 (-0.45) | 1.39 (1.77) | 1.10 (3.33) | 100 |
| | | Two-stage II | -0.94 (-0.86) | 1.42 (1.74) | 1.29 (3.77) | 100 |
| | | Two-stage III | -1.11 (-0.76) | 1.37 (2.45) | 1.46 (6.61) | 94.3 |
| | | GMM | -1.67 (-1.63) | 1.18 (1.63) | 1.75 (5.34) | 56.3 |
| 0.25 | (-2, -2.00) | Exact I | -0.31 (-0.15) | 0.83 (1.12) | 0.70 (1.27) | 79.1 |
| | | Approximate | 0.19 (0.26) | 0.99 (1.24) | 0.71 (1.60) | 100 |
| | | Adjusted | 0.19 (0.26) | 0.99 (1.24) | 0.71 (1.60) | 100 |
| | | Logist. reg. | -1.95 (-1.95) | 0.14 (0.15) | 1.95 (3.81) | 100 |
| | | Two-stage I | 0.13 (0.19) | 0.96 (1.20) | 0.68 (1.48) | 100 |
| | | Two-stage II | 0.13 (0.19) | 0.96 (1.20) | 0.69 (1.20) | 100 |
| | | Two-stage III | 0.40 (1.36) | 1.29 (4.91) | 1.03 (25.94) | 88.0 |
| | | GMM | -0.14 (0.02) | 1.12 (1.38) | 0.84 (1.90) | 49.1 |
| 0.25 | (0, -2.58) | Exact I | -0.27 (-0.23) | 0.99 (1.54) | 0.72 (2.41) | 84.5 |
| | | Approximate | -0.52 (-0.49) | 0.87 (1.01) | 0.73 (1.27) | 100 |
| | | Adjusted | -0.61 (-0.60) | 0.97 (1.13) | 0.85 (1.64) | 100 |
| | | Logist. reg. | 0.01 (0.02) | 0.20 (0.22) | 0.15 (0.05) | 100 |
| | | Two-stage I | -0.20 (-0.19) | 0.94 (1.10) | 0.70 (1.24) | 100 |
| | | Two-stage II | -0.37 (-0.34) | 0.95 (1.11) | 0.74 (1.36) | 100 |
| | | Two-stage III | -0.45 (-0.05) | 1.02 (2.39) | 0.92 (5.74) | 97.1 |
| | | GMM | -0.58 (-0.35) | 1.01 (1.65) | 0.96 (2.84) | 75.9 |
| 0.50 | (-2, -1.03) | Exact I | -0.04 (0.07) | 0.64 (0.88) | 0.50 (0.79) | 96.0 |
| | | Approximate | 0.38 (0.49) | 0.87 (1.14) | 0.67 (1.54) | 100 |
| | | Adjusted | 0.38 (0.49) | 0.87 (1.14) | 0.66 (1.54) | 100 |
| | | Logist. reg. | -1.93 (-1.94) | 0.12 (0.13) | 1.93 (3.77) | 100 |
| | | Two-stage I | 0.45 (0.57) | 0.90 (1.17) | 0.70 (1.70) | 100 |
| | | Two-stage II | 0.45 (0.57) | 0.90 (1.17) | 0.70 (1.70) | 100 |
| | | Two-stage III | 0.95 (2.14) | 1.30 (10.36) | 1.16 (111.83) | 81.9 |
| | | GMM | 0.46 (0.71) | 0.97 (1.85) | 0.79 (3.93) | 80.2 |
| 0.50 | (0, -1.27) | Exact I | -0.0006 (0.02) | 0.85 (1.00) | 0.60 (1.00) | 97.1 |
| | | Approximate | -0.15 (-0.15) | 0.70 (0.80) | 0.53 (0.66) | 100 |
| | | Adjusted | -0.49 (-0.50) | 0.87 (0.99) | 0.70 (1.23) | 100 |
| | | Logist. reg. | -0.006 (0.002) | 0.10 (0.15) | 0.02 (0.02) | 100 |
| | | Two-stage I | 0.06 (0.05) | 0.85 (0.98) | 0.63 (0.96) | 100 |
| | | Two-stage II | 0.23 (0.22) | 0.85 (0.95) | 0.63 (0.97) | 100 |
| | | Two-stage III | 0.26 (0.68) | 1.11 (2.00) | 0.85 (4.48) | 96.8 |
| | | GMM | 0.009 (0.16) | 0.90 (1.41) | 0.66 (2.02) | 96.5 |

equal to $-1$ or 0 (we chose smaller effect sizes because $X_i$ was now more variable than before). This setting is advantageous to the Two-stage estimators which were developed, assuming a normally distributed exposure. The results from these simulation studies are summarized in Tables 3.6 and 3.7. They reveal qualitatively similar conclusions to before, apart from demonstrating a much better performance of the Adjusted IV-estimator. Results for Exact IV-estimator II are not reported because they were very comparable with those for Exact IV-estimator I and required long convergence times for completion.

Table 3.6: *Simulation results when exposure $X$ is continuous, $\psi^* = 1$ and $\rho_{xz} = 0.26$.*

| | | | **Bias** | **SE** | | |
|---|---|---|---|---|---|---|
| E(**Y**) | ($\eta^*$, $\omega^*$) | **Methods** | **Median (Mean)** | **MCD (Sample)** | **MAD (MSE)** | % **conv.** |
| 0.05 | (-1, -3.28) | Exact I | -0.04 (0.14) | 0.57 (0.82) | 0.49 (0.69) | 99.4 |
| | | Approximate | -0.02 (0.01) | 0.51 (0.58) | 0.37 (0.33) | 100 |
| | | Adjusted | -0.01 (0.01) | 0.51 (0.58) | 0.37 (0.33) | 100 |
| | | Logist. reg. | -0.93 (-0.93) | 0.12 (0.13) | 0.93 (0.88) | 100 |
| | | Two-stage I | -0.25 (-0.23) | 0.39 (0.43) | 0.33 (0.24) | 100 |
| | | Two-stage II | -0.25 (-0.23) | 0.39 (0.43) | 0.34 (0.24) | 100 |
| | | Two-stage III | -0.17 (0.07) | 0.48 (1.28) | 0.45 (1.64) | 97.9 |
| | | GMM | 0.03 (0.81) | 0.58 (15.54) | 0.44 (242.12) | 95.9 |
| 0.05 | (0, -4.15) | Exact I | -0.05 (0.17) | 0.52 (0.94) | 0.44 (0.92) | 95.4 |
| | | Approximate | -0.08 (-0.05) | 0.48 (0.54) | 0.36(0.29) | 100 |
| | | Adjusted | -0.02 (0.01) | 0.50 (0.56) | 0.38 (0.32) | 100 |
| | | Logist. reg. | 0.008 (0.01) | 0.14 (0.15) | 0.11 (0.02) | 100 |
| | | Two-stage I | -0.22 (-0.21) | 0.40 (0.44) | 0.34 (0.24) | 100 |
| | | Two-stage II | -0.22 (-0.20) | 0.40 (0.45) | 0.34 (0.24) | 100 |
| | | Two-stage III | -0.24 (-0.08) | 0.42 (0.82) | 0.43 (0.68) | 98.2 |
| | | GMM | -0.05 (0.24) | 0.53 (3.37) | 0.43 (11.42) | 97.9 |
| 0.25 | (-1, -1.58) | Exact I | -0.004 (0.04) | 0.29 (0.35) | 0.21 (0.12) | 100 |
| | | Approximate | 0.008 (0.02) | 0.26 (0.30) | 0.20 (0.09) | 100 |
| | | Adjusted | 0.009 (0.02) | 0.27 (0.30) | 0.20 (0.09) | 100 |
| | | Logist. reg. | -0.93 (-0.93) | 0.06 (0.06) | 0.93 (0.86) | 100 |
| | | Two-stage I | -0.02 (-0.007) | 0.26 (0.29) | 0.19 (0.08) | 100 |
| | | Two-stage II | -0.02 (-0.007) | 0.26 (0.29) | 0.19 (0.08) | 100 |
| | | Two-stage III | 0.21 (0.47) | 0.40 (1.30) | 0.31 (1.90) | 98.5 |
| | | GMM | 0.26 (0.90) | 0.43 (7.94) | 0.36 (63.90) | 97.4 |
| 0.25 | (0, -2.09) | Exact I | 0.009 (0.05) | 0.31 (0.39) | 0.23 (0.16) | 100 |
| | | Approximate | -0.15 (-0.14) | 0.25 (0.26) | 0.21 (0.09) | 100 |
| | | Adjusted | -0.001 (0.008) | 0.29 (0.31) | 0.21 (0.09) | 100 |
| | | Logist. reg. | -0.002 (0.003) | 0.08 (0.90) | 0.06 (0.008) | 100 |
| | | Two-stage I | -0.05 (-0.04) | 0.28 (0.29) | 0.21 (0.09) | 100 |
| | | Two-stage II | -0.04 (-0.04) | 0.28 (0.29) | 0.21 (0.09) | 100 |
| | | Two-stage III | -0.04 (0.04) | 0.34 (0.53) | 0.29 (0.28) | 99.8 |
| | | GMM | -0.006 (0.11) | 0.37 (0.71) | 0.29 (0.51) | 99.8 |
| 0.50 | (-1, -0.64) | Exact I | 0.01 (0.02) | 0.24 (0.28) | 0.18 (0.08) | 100 |
| | | Approximate | 0.01 (0.01) | 0.23 (0.26) | 0.17 (0.07) | 100 |
| | | Adjusted | 0.01 (0.01) | 0.23 (0.26) | 0.17 (0.07) | 100 |
| | | Logist. reg. | -0.93 (-0.93) | 0.06 (0.06) | 0.93 (0.87) | 100 |
| | | Two-stage I | 0.04 (0.04) | 0.24 (0.27) | 0.18 (0.07) | 100 |
| | | Two-stage II | 0.04 (0.04) | 0.24 (0.27) | 0.18 (0.07) | 100 |
| | | Two-stage III | 0.32 (0.55) | 0.44 (1.22) | 0.38 (1.79) | 97.9 |
| | | GMM | 0.36 (0.87) | 0.47 (2.00) | 0.41 (4.41) | 97.2 |
| 0.50 | (0, -0.77) | Exact I | 0.009 (0.02) | 0.24 (0.27) | 0.18 (0.07) | 100 |
| | | Approximate | -0.16 (-0.16) | 0.20 (0.22) | 0.19 (0.07) | 100 |
| | | Adjusted | 0.01 (0.01) | 0.24 (0.26) | 0.18 (0.07) | 100 |
| | | Logist. reg. | -0.0004 (0.005) | 0.08 (0.08) | 0.05 (0.007) | 100 |
| | | Two-stage I | 0.004 (0.004) | 0.24 (0.26) | 0.18 (0.07) | 100 |
| | | Two-stage II | 0.005 (0.005) | 0.24 (0.26) | 0.17 (0.07) | 100 |
| | | Two-stage III | 0.01 (0.08) | 0.30 (0.54) | 0.24 (0.30) | 100 |
| | | GMM | 0.01 (0.09) | 0.31 (0.55) | 0.25 (0.31) | 99.9 |

Table 3.7: *Simulation results when exposure $X$ is continuous, $\psi^* = 0$ and $\rho_{xz} = 0.26$.*

| | | | **Bias** | **SE** | | |
|---|---|---|---|---|---|---|
| E(**Y**) | ($\eta^*$, $\omega^*$) | **Methods** | **Median (Mean)** | **MCD (Sample)** | **MAD (MSE)** | % **conv.** |
| 0.05 | (-1, -2.9) | Exact I | 0.02 (0.01) | 0.51 (0.65) | 0.37 (0.42) | 96.6 |
| | | Approximate | -0.01 (-0.01) | 0.47 (0.54) | 0.35 (0.30) | 100 |
| | | Adjusted | -0.02 (-0.01) | 0.51 (0.57) | 0.38 (0.33) | 100 |
| | | Logist. reg. | -0.93 (-0.93) | 0.14 (0.15) | 0.93 (0.90) | 100 |
| | | Two-stage I | -0.009 (-0.01) | 0.40 (0.45) | 0.30 (0.21) | 100 |
| | | Two-stage II | -0.01 (-0.01) | 0.41 (0.46) | 0.30 (0.21) | 100 |
| | | Two-stage III | -0.009 (-0.01) | 0.37 (0.50) | 0.27 (0.25) | 100 |
| | | GMM | -0.01 (-0.01) | 0.47 (0.59) | 0.36 (0.35) | 100 |
| 0.05 | (0, -2.9) | Exact I | 0.008 (0.04) | 0.45 (0.62) | 0.33 (0.39) | 98.1 |
| | | Approximate | 0.004 (-0.01) | 0.45 (0.52) | 0.32 (0.27) | 100 |
| | | Adjusted | 0.004 (-0.01) | 0.45 (0.52) | 0.32 (0.27) | 100 |
| | | Logist. reg. | 0.003 (0.003) | 0.13 (0.14) | 0.09 (0.02) | 100 |
| | | Two-stage I | 0.005 (-0.009) | 0.34 (0.39) | 0.25 (0.16) | 100 |
| | | Two-stage II | 0.003 (-0.009) | 0.34 (0.39) | 0.24 (0.16) | 100 |
| | | Two-stage III | 0.003 (-0.005) | 0.33 (0.48) | 0.25 (0.23) | 100 |
| | | GMM | 0.002 (-0.01) | 0.45 (0.59) | 0.32 (0.35) | 99.8 |
| 0.25 | (-1, -1.09) | Exact I | 0.01 (0.02) | 0.29 (0.33) | 0.21 (0.11) | 100 |
| | | Approximate | 0.01 (0.01) | 0.24 (0.27) | 0.18 (0.07) | 100 |
| | | Adjusted | 0.01 (0.02) | 0.28 (0.32) | 0.21 (0.10) | 100 |
| | | Logist. reg. | -0.92 (-0.92) | 0.08 (0.09) | 0.92 (0.86) | 100 |
| | | Two-stage I | 0.01 (0.02) | 0.27 (0.30) | 0.20 (0.09) | 100 |
| | | Two-stage II | 0.01 (0.02) | 0.27 (0.31) | 0.20 (0.09) | 100 |
| | | Two-stage III | 0.01 (0.02) | 0.24 (0.28) | 0.18 (0.08) | 100 |
| | | GMM | 0.01 (0.02) | 0.24 (0.29) | 0.18 (0.08) | 100 |
| 0.25 | (0, -1.09) | Exact I | 0.007 (0.004) | 0.26 (0.28) | 0.18 (0.08) | 100 |
| | | Approximate | 0.006 (0.001) | 0.25 (0.28) | 0.18 (0.08) | 100 |
| | | Adjusted | 0.006 (0.001) | 0.25 (0.28) | 0.18 (0.08) | 100 |
| | | Logist. reg. | 0.002 (0.003) | 0.06 (0.07) | 0.05 (0.005) | 100 |
| | | Two-stage I | 0.007 (0.001) | 0.25 (0.27) | 0.17 (0.07) | 100 |
| | | Two-stage II | 0.006 (0.001) | 0.25 (0.27) | 0.17 (0.07) | 100 |
| | | Two-stage III | 0.006 (0.001) | 0.25 (0.28) | 0.17 (0.08) | 100 |
| | | GMM | 0.006 (0.001) | 0.25 (0.29) | 0.18 (0.08) | 100 |
| 0.50 | (-1, 0) | Exact I | 0.01 (0.01) | 0.27 (0.30) | 0.19 (0.09) | 100 |
| | | Approximate | 0.008 (0.01) | 0.22 (0.24) | 0.16 (0.06) | 100 |
| | | Adjusted | 0.005 (0.01) | 0.27 (0.29) | 0.19 (0.08) | 100 |
| | | Logist. reg. | -0.92 (-0.92) | 0.07 (0.08) | 0.92 (0.85) | 100 |
| | | Two-stage I | 0.009 (0.01) | 0.27 (0.29) | 0.19 (0.08) | 100 |
| | | Two-stage II | 0.01 (0.01) | 0.27 (0.29) | 0.19 (0.08) | 100 |
| | | Two-stage III | 0.009 (0.01) | 0.23 (0.26) | 0.16 (0.07) | 100 |
| | | GMM | 0.008 (0.01) | 0.23 (0.25) | 0.16 (0.06) | 100 |
| 0.50 | (0 , 0) | Exact I | 0.004 (0.001) | 0.23 (0.25) | 0.16 (0.06) | 100 |
| | | Approximate | 0.004 (0.002) | 0.23 (0.25) | 0.16 (0.06) | 100 |
| | | Adjusted | 0.004 (0.002) | 0.23 (0.25) | 0.16 (0.06) | 100 |
| | | Logist. reg. | 0.003 (0.002) | 0.06 (0.06) | 0.04 (0.004) | 100 |
| | | Two-stage I | 0.004 (0.002) | 0.23 (0.26) | 0.17 (0.07) | 100 |
| | | Two-stage II | 0.004 (0.002) | 0.23 (0.26) | 0.17 (0.07) | 100 |
| | | Two-stage III | 0.004 (0.002) | 0.23 (0.27) | 0.17 (0.07) | 100 |
| | | GMM | 0.004 (0.002) | 0.23 (0.26) | 0.16 (0.07) | 100 |

# 5 Discussion

In this article, we have given an expository review of IV-estimators for the causal odds ratio. We have focused on exact estimators as well as a number of popular approximate ones, without being exhaustive; in particular, we have omitted estimators based on principal stratification (e.g. Abadie, 2003; Ten Have et al., 2003) as this approach does not allow a flexible treatment of continuous exposures and is rather artificial in the context of Mendelian randomization studies (Didelez, Meng and Sheehan, 2008). Our results show that the concerns of Robins and Rotnitzky (2004) about incongeniality of the model of Vansteelandt and Goetghebeur (2003) can be overruled by leaving the main effect of the IV in their association model unrestricted. This is useful because the approach of Vansteelandt and Goetghebeur (2003) yields computationally simpler estimators through the following three steps: (a) fit a standard association model for the outcome in function of exposure and IV; (b) combine this with the causal model to obtain predicted values of the exposure-free outcome; and (c) choose the causal odds ratio such that these predictions become independent of the IV. For general IVs, leaving its main effect unrestricted requires the use of generalized additive association models. The performance of the resulting estimator remains to be studied. Our simulation studies complement recent studies by Didelez, Meng and Sheehan (2008), Palmer et al. (2008) and Rassen et al. (2009), but include results on the bias and efficiency of exact IV-estimators. They reveal that these 'exact' estimators tend to outperform 'approximate' estimators of the causal odds ratio that are commonly used in the literature on Mendelian randomization. Of all considered estimators, the Exact IV-estimators are the only ones which are asymptotically unbiased, although they may have an important finite-sample bias when there is limited information (e.g. due to low prevalence, weak IV, small sample size, ...). The Approximate IV-estimator tended to outperform standard logistic regression when there was confounding of a sufficient magnitude, and was doing especially well at the causal null hypothesis. This estimator has a number of attractions over the Exact IV-estimators in that it can be used for the analysis of case-control data (Smith et al., 2005) and lends itself particularly well to meta-analyses based on summary measures (Minelli et al., 2004). In addition, it can be extended to the analysis of time-varying exposures. Indeed, consider the following logistic structural nested mean model

$$\begin{aligned}&\text{logit}E(Y_{t(\overline{x}_s 0)}|\overline{X}_s = \overline{x}_s, Z)\\ &\qquad -\text{logit}E(Y_{t(\overline{x}_{s-1} 0)}|\overline{X}_s = \overline{x}_s, Z) = \psi^* I(s=t)x_t + \gamma^* I(s<t)x_s\end{aligned}$$

where $Y_{t(\overline{x}_s 0)}$ denotes the counterfactual outcome that would be observed at time $t$ if, possibly contrary to fact, the exposure history $\overline{X}_t = (X_1, ..., X_t)$ equalled $(\overline{x}_s, 0, ..., 0)$.

This model allows for the short term effect $\psi^*$ to differ from the long term effect $\gamma^*$. Using similar approximations as in Section 2.4, it can be shown that this model implies

$$\text{logit}E(Y_t|Z) = \omega_t^* + \psi^* E(X_t|Z) + \gamma^* \sum_{s=1}^{t-1} E(X_s|Z)$$

where $Y_t$ is the observed outcome at time $t > 0$. Having obtained estimates of $E(X_s|Z)$ for $s > 0$, this model can be fitted - and thus estimates of $\psi^*$ and $\gamma^*$ can be obtained - using standard software for marginal models. The resulting estimators continue to share the local robustness property of being consistent at the causal null hypothesis. It remains to be evaluated to what extent the approximation errors propagate with time and thus how prone to bias these estimators are away from the causal null hypothesis. It additionally remains to be studied whether similar adjustments as for the Adjusted IV-Estimator (possibly including the previously suggested corrections for measurement error) can remedy some of the bias of the resulting estimator.

With the exception of GMM-estimators, our focus in this article has been on estimation of the exposure effect conditional on the observed exposure, as defined in (3.1). In linear structural mean models (Robins, 1994), the assumption that the treatment effect is not modified by the instrument, i.e. that

$$E(Y - Y_0|X, Z) = \psi^* X$$

does not depend on $Z$, implies that marginal and conditional effects are the same. This is seen because the above model implies the same observed data restriction, namely $E(Y - \psi^* X|Z) = E(Y - \psi^* X)$, as model $E(Y - Y_1|X, Z) = \phi^*(X - 1)$, thus indicating that

$$E(Y_1 - Y_0|X = 1, Z) \equiv \psi^* = \phi^* \equiv E(Y_1 - Y_0|X = 0, Z)$$

and, consequently, that $\psi^* = E(Y_1 - Y_0)$. The same is no longer true in logistic structural mean models, where additional assumptions are required to infer marginal causal effects.

Adjustment for baseline covariates (more generally, covariates which are not causally affected by exposure, outcome or IV) was not discussed in this article because it is not commonly considered in biostatistical and epidemiological applications. This is justified by the fact that covariate adjustment is only needed when there are measured confounders of the association between IV and outcome, when there is a specific interest in assessing effect modification or when an boost in efficiency is anticipated. Covariate adjustment is easily realized for all considered IV-estimators by additionally including these in all considered regression models. For the Exact IV-estimators,

this requires testing whether the predicted counterfactual outcome $Y_0$ is independent of the IV, *conditional* on baseline covariates. When there is a continuous baseline covariate or multiple discrete covariates, then the model of Vansteelandt and Goetghebeur (2003) is no longer guaranteed to yield a congenial parameterization, unlike the model of Robins and Rotnitzky (2004). The similarity of the estimating functions indicates that, nevertheless, similar estimates would typically be obtained with both approaches.

# Appendix 3.A: Local robustness and Incongenial models

## Local robustness

When $\psi^* = 0$, then equation (3.7) becomes $\sum_{i=1}^{n} \left( Z_i - \frac{\sum_{j=1}^{n} Z_j}{n} \right) \text{expit}\{m(X_i, Z_i; \hat{\beta})\}$. Suppose now that the association model includes an intercept and main effect in $Z_i$, and that $\hat{\beta}$ is the standard maximum likelihood estimator of $\beta^*$. We then show that equation (3.7) equals $\sum_{i=1}^{n} \left( Z_i - \frac{\sum_{j=1}^{n} Z_j}{n} \right) Y_i$, which has mean zero at $\psi^* = 0$, even under model misspecification. That this equality is true follows because $\hat{\beta}$ satisfies the following score equations:

$$0 = \sum_{i=1}^{n} \begin{pmatrix} 1 \\ Z_i \end{pmatrix} \left[ Y_i - \text{expit}\{m(X_i, Z_i; \hat{\beta})\} \right]$$

from which $\sum_{i=1}^{n} Z_i Y_i = \sum_{i=1}^{n} Z_i \text{expit}\{m(X_i, Z_i; \hat{\beta})\}$ and $\sum_{i=1}^{n} \frac{\sum_{j=1}^{n} Z_j}{n} Y_i = \sum_{i=1}^{n} \frac{\sum_{j=1}^{n} Z_j}{n}$ $\times \text{expit}\{m(X_i, Z_i; \hat{\beta})\}$.

## Incongenial models

It follows from the parameterization of Robins and Rotnitzky (2004) that the logistic structural mean model (3.2) is congenial with an association model of the form

$$P(Y_i = 1 | X_i, Z_i) = \text{expit}\{\psi^* X_i + q(X_i, Z_i) + v(Z_i)\}$$

where $q(X_i, Z_i)$ is an arbitrary function of $(X_i, Z_i)$ satisfying $q(0, Z_i) = 0$ for all $Z_i$ and where $v(Z_i)$ solves

$$\omega \quad = \quad \int \text{expit}\{q(X_i = x, Z_i) + v(Z_i)\}P(X_i = x|Z_i)dx$$

for some value $\omega$. Consider now an association model of the form

$$P(Y_i = 1|X_i, Z_i; \beta^*) = \text{expit}\{\beta_1^* X_i + q^*(X_i, Z_i; \beta_2^*) + v^*(Z_i; \beta_3^*)\}$$

with $v^*(Z_i; \beta_3)$ and $q^*(X_i, Z_i; \beta_2)$ arbitrary functions of $(Z_i, \beta_3)$ and of $(X_i, Z_i)$, satisfying $q^*(0, Z_i; \beta_2) = 0$ for all $Z_i$ and $\beta_2$, respectively. It then follows from the stated results of Robins and Rotnitzky (2004) that this association model is congenial with the logistic structural mean model (3.2) for every choice of model $q^*(X_i, Z_i; \beta_2)$ when no restrictions are imposed on the function $v^*(Z_i; \beta_3)$, e.g. when $Z_i$ is dichotomous and $v^*(Z_i; \beta_3)$ is chosen to be of the form $\beta_{30} + \beta_{31} Z_i$ for unknown parameters $\beta_{30}, \beta_{31}$.

## Two-stage estimator

In this section, we explain how to derive $\text{E}(Y_{i0}|Z_i)$ under models (3.12) and (3.13). Note that

$$\text{E}(Y_{i0}|Z_i, X_i) \quad = \quad P(U_i \leq \theta_0^* + \theta_1^* X_i + \theta_2^* Z_i - \phi^* X_i)$$

where $U_i$ is a standard normally distributed variate, independent of $(Z_i, X_i)$. Averaging over the exposure, conditional on $Z_i$, then yields

$$\text{E}(Y_{i0}|Z_i) = \quad \int_{-\infty}^{\infty} P(U_i + (\phi^* - \theta_1^*)X_i \leq \theta_0^* + \theta_2^* Z_i) dF(X_i|Z_i)$$

where $F(X_i|Z_i)$ refers to the conditional distribution of $X_i$, given $Z_i$. Define $U_i^* = U_i + (\phi^* - \theta_1^*)X_i$. Then, assuming that $X_i$ is normally distributed with mean $\alpha_0^* + \alpha_1^* Z_i$ and constant variance $\sigma^{2*}$, conditional on $Z_i$, $U_i^*$ has a normal distribution with mean $\mu_{u^*} = (\phi^* - \theta_1^*)(\alpha_0 + \alpha_1 Z_i)$ and variance $\sigma_{u^*}^2 = 1 + (\phi^* - \theta_1^*)^2 \sigma^2$. Then

$$\text{E}(Y_{i0}|Z_i) \quad = \quad \int_{-\infty}^{\infty} \int_{-\infty}^{\theta_0^* + \theta_2^* Z_i} dF(U_i^*, X_i|Z_i) = \Phi\{(\theta_0^* + \theta_2^* Z_i - \mu_{u^*})/\sigma_{u^*}\}.$$

The conditional mean $\text{E}(Y_i|Z_i)$ can be derived using similar arguments.

# Chapter 4

# Correcting Instrumental Variables Estimators for Systematic Measurement Error


## Summary

Instrumental variables (IV) estimators are well established to correct for measurement error on exposure in a broad range of fields. In a distinct prominent stream of research IV's are becoming increasingly popular for estimating causal effects of exposure on outcome since they allow for unmeasured confounders which are hard to avoid. Because many causal questions emerge from data which suffer severe measurement error problems, we combine both IV approaches in this article to correct IV-based causal effect estimators in linear (structural mean) models for possibly systematic measurement error on the exposure. The estimators rely on the presence of a baseline measurement which is associated with the observed exposure and known not to modify the target effect. Simulation studies and the analysis of a small blood pressure reduction trial ($n = 105$) with treatment noncompliance confirm the adequate performance of our estimators in finite samples. Our results also demonstrate that incorporating limited prior knowledge about a weakly identified parameter (such as the error mean) in a frequentist analysis can yield substantial improvements.

# 1 Introduction

Instrumental variables (IV) methods have a long tradition in economics and econometrics, where they are used in connection with structural equation models. They have more recently entered the medical, epidemiological and biostatistical literature (for reviews, see e.g. Greenland, 2000; Martens et al., 2006). To estimate the average causal effect of an exposure on an outcome in the presence of unmeasured confounders, these methods rely on so-called instrumental variables. These are variables which (i) are associated with the exposure; (ii) have no direct effect on the outcome; and (iii) do not share common causes with the outcome (Hernán and Robins, 2006). Instrumental variables arise naturally in double-blind randomized trials with treatment noncompliance because randomization (i.e. the instrument) is associated with received treatment (i.e. exposure), often does not affect outcome other than through received treatment and shares no common causes with outcome by virtue of randomization. They are hence frequently used to adjust for treatment noncompliance in randomized experiments (see e.g. Goetghebeur and Vansteelandt, 2005, for a review) and for the analysis of randomized encouragement designs (Ten Have et al., 2004). At the same time, they are becoming increasingly popular in observational settings where the conditions for an instrumental variable are harder to justify. In genetics, for instance, the random assortment of genes transferred from parents to offspring - called 'Mendelian randomization' - resembles the use of randomization in experiments and is therefore a natural instrumental variable for estimating the effect of genetically affected exposures on a given trait (Didelez and Sheehan, 2007). Casas et al. (2005) use this idea to assess the influence of plasma homocysteine level on the risk of stroke with homozygosity at a specific allele as an instrumental variable. In most observational studies no real or natural randomization is present, in which case the availability of an instrumental variable must be assessed on theoretical grounds. For instance, Leigh and Schembri (2004) use the cigarette price per region as an instrumental variable to estimate the effect of smoking on health, assuming that the price of cigarettes may only impact health by mediating exposure to cigarette smoke.

With the increasing popularity of IV methods in causal inference comes the growing concern for their performance under common complications, such as misclassification or measurement error on exposure. Indeed, in the context of noncompliance adjustment in clinical trials (Dunn, 1999; Goetghebeur and Vansteelandt, 2005) for instance, simple measures of compliance with drug therapy, such as pill counts, are notorious for overestimating the amount of drug actually taken (Urquhart and De Klerk, 1998). HIV prevention studies tend to rely on self reported measures of sexual activity and accompanying preventive action, including use of condoms or microbicide gels, which are subject to 'pleasing bias' (Van Damme et al., 2005). Many other

exposure measures are popular even though they are bias prone.

Random measurement error on exposure is not alarming for IV estimators in linear (structural mean) models (Goetghebeur and Vansteelandt, 2005). These estimators continue to be asymptotically unbiased when random measurement error is ignored, with at most a slight loss of efficiency. When measurement error is systematic, tests of the causal null hypothesis of no effect remain valid, but effect estimates may become biased. Because systematic error is a real concern in many practical settings (e.g. overreporting of drug compliance, underreporting of alcohol use, ...), our goal in this article is to investigate how IV estimators for the parameters in linear (structural mean) models may be adjusted for systematic measurement error. Goetghebeur and Vansteelandt (2005) show how this can be done when the average size of the error is known. This allows for sensitivity analyses, but leaves open the question of how to estimate the average size of the measurement error and subsequently correct for it. Because of identifiability problems, the latter can only be realized when extraneous information is available. One common source of information is an IV for the measurement error (Buzas and Stefanski, 1996; Carroll et al., 2004, 2006). In contrast to the original IV used for confounder adjustment, this is a (pre-exposure) surrogate for the observed exposure (in the sense that it is correlated with exposure), which is assumed not to modify the exposure effect of interest. Our interest in such variables stems from the fact that we can identify settings where such variables exist and that other sources of information on the measurement error, such as repeated measurements or validation samples, are typically not available in large classes of problems (e.g. noncompliance adjustment).

In the next section, we build on ideas from linear regression models with error in the covariates (Carroll et al., 2006) to show how an IV for the measurement error can help correct IV-based causal effect estimators for systematic error under linear structural mean models (Goetghebeur and Lapp, 1997; Robins, 1994). In Section 2.3, we diagnose poor performance of the error-adjusted estimator in small to moderate sample sizes as compared to the standard estimator which ignores measurement error. We show in Section 3 that this is due to the average magnitude of the error being weakly identified at effects close to zero. In Section 3, we accommodate this by imposing liberal bounds on the magnitude of the average error. This leads to reliable estimators for the causal effect of observed exposure with good performance in finite samples. The latter is confirmed through the analysis of a placebo-controlled hypertension trial in Section 4 and through simulation studies in Section 3.2. Our results reveal how the incorporation of prior information (in the form of bounds on weakly identified nuisance parameters) in a frequentist analysis, can recover considerable precision for the target parameter.

# 2 Adjusting for measurement error

## 2.1 Assumptions

We consider data on a scalar exposure $Z_i$, a scalar outcome $Y_i$ and possibly a vector of baseline (i.e. pre-exposure) covariates $\mathbf{X}_i$ drawn from independent subjects $i = 1, ..., n$, to study the average effect of exposure $Z_i$ on outcome $Y_i$. We define this effect as an expected contrast

$$\mathrm{E}(Y_i - Y_{i0}|Z_i, \mathbf{X}_i), \tag{4.1}$$

between observed outcomes $Y_i$ and potential exposure-free outcomes $Y_{i0}$ (Rubin, 1978). The latter indicates a reference response which would have been measured for subject $i$ if all conditions had been the same as in the current study, but no exposure had been received; that is, if the assigned experimental treatment contained no active dose. Because true exposure $Z_i$ is imprecisely measured, the observed exposure level $W_i$ for subject $i$ may differ from the actual exposure level $Z_i$, which is unobserved.

Since $Y_{i0}$ and $Z_i$ are not generally observed, identification of the causal effect (4.1) requires assumptions.

*Assumption A1 (Causal IV assumption):* $R_i$ is a *causal instrumental variable (IV-C)* for inferring the causal effect of $Z_i$ on $Y_i$; that is, $R_i$ is conditionally dependent on $Z_i$, given $\mathbf{X}_i$, and satisfies the following assumptions:

1. exclusion restriction (Angrist, Imbens and Rubin, 1996): $R_i$ has no direct effect on the outcome (only an indirect effect via the exposure is possible). That is, with $Y_{i0r}$ the potential outcome that we would have observed for subject $i$ if $(R_i, Z_i)$ were set to $(r, 0)$, we assume that $Y_{i0r} = Y_{i0}$ for all values of $r$ in the support of $R_i$.

2. randomization assumption: within strata of baseline covariates $\mathbf{X}_i$, $\mathrm{E}(Y_{i0}|\mathbf{X}_i, R_i) = \mathrm{E}(Y_{i0}|\mathbf{X}_i)$.

In double-blind randomized trials of an asymptomatic disease, one expects these assumptions to hold for randomization $R_i$ since patients and physicians are unaware of the assigned treatment (Robins, 1994).

*Assumption A2 (Consistency assumption):* to link exposure-free outcomes to observed outcomes, we assume that $Y_i = Y_{i0}$ for subjects with $Z_i = 0$.

*Assumption A3 (Model assumption):* the expected causal effect (4.1) follows the linear structural mean model (Robins, 1994; Goetghebeur and Lapp, 1997)

$$\mathrm{E}(Y_i - Y_{i0}|Z_i, \mathbf{X}_i, R_i) = \gamma(\mathbf{X}_i; \boldsymbol{\psi}^*)Z_i \tag{4.2}$$

where $\gamma(\mathbf{X}_i;\boldsymbol{\psi})$ is a known function smooth in $\boldsymbol{\psi}$, satisfying $\gamma(\mathbf{X}_i;\mathbf{0})=0$, and where $\boldsymbol{\psi}^*$ is an unknown finite-dimensional parameter. For instance, in placebo-controlled randomized experiments with $R_i=1$ for subjects randomized to the experimental arm and $R_i=0$ for placebo control and with $Z_i$ denoting exposure to the experimental treatment, we may choose

$$\mathrm{E}(Y_i - Y_{i0}|Z_i,\mathbf{X}_i,R_i) = \psi^* Z_i. \tag{4.3}$$

Here, $\psi^*$ expresses the expected change in outcome when those exposed to $Z_i=1$ would have their exposure set to zero. When treatment effects are potentially modified by pre-treatment covariates, one may add covariate-exposure interactions, as in

$$\mathrm{E}(Y_i - Y_{i0}|Z_i,\mathbf{X}_i,R_i) = (\psi_1^* + \psi_2^{*'}\mathbf{X}_i)Z_i.$$

Here, $\boldsymbol{\psi}_2^*$ defines the change in the average effect of unit exposure per unit increase in $\mathbf{X}_i$.

Note that we restrict our development to models (4.2) which postulate the causal effect to be linear in the exposure. This is a standard restriction in the literature on IV-estimation and on two-stage-least-squares estimation of causal effects (Hernán and Robins, 2006) because linear structural mean models with nonlinear exposure effects suffer from identification problems, even in the absence of measurement error (Vansteelandt and Goetghebeur, 2005). For similar reasons, no effect modification by the IV-C is allowed. Note, however, that model (4.2) will often give a reasonable approximation, even for nonlinear causal effects.

*Assumption A4 (Measurement error IV assumption):* Given the difficulty in obtaining information about measurement error characteristics, we introduce an *instrumental variable for the measurement error (IV-M).* In contrast to an IV-C which satisfies Assumption A1, we define this as a surrogate $\mathbf{T}_i \subseteq \mathbf{X}_i$ for the observed exposure (in the sense that is it is conditionally associated with $W_i$, given $(\mathbf{S}_i,R_i)$, where $\mathbf{S}_i$ is such that $\mathbf{X}_i \equiv (\mathbf{S}_i,\mathbf{T}_i)$), which is measured prior to exposure and is such that it does not modify the causal effect of received exposure on the outcome, i.e. such that

$$\mathrm{E}(Y_i - Y_{i0}|Z_i,\mathbf{X}_i,R_i) = \mathrm{E}(Y_i - Y_{i0}|Z_i,\mathbf{S}_i,R_i). \tag{4.4}$$

We thus assume that $\gamma(\mathbf{X}_i;\boldsymbol{\psi})$ in (4.2) does not involve $\mathbf{T}_i$. With a slight abuse of notation, we will denote it $\gamma(\mathbf{S}_i;\boldsymbol{\psi})$. Importantly, note that the IV-M $\mathbf{T}_i$ differs from and satisfies different assumptions than the IV-C $R_i$, which satisfies Assumption A1. The former instrumental variable will be used to correct for systematic measurement error, the latter to infer a causal effect of $Z$ on $Y$.

The use of no-interaction assumptions such as (4.4) is increasingly common in causal inference, in particular in the context of IV-estimation. For instance, Ten

Have et al. (2007), Joffe, Small and Hsu (2007) and Albert (2008) use similar no-interaction assumptions to infer direct causal effects. Vansteelandt and Goetghebeur (2004) and Fischer and Goetghebeur (2006) rely on no-interaction assumptions for assessing effect modification by treatment-free responses. In this study, the interest in Assumption A4 is motivated by the fact that other sources of information on the measurement error, such as repeated measurements or validation samples, are typically not available in large classes of problems (e.g. noncompliance adjustment), and by the fact that we can identify settings where the assumption is reasonable. For instance, in randomized clinical trials, one source of an IV for the measurement error on treatment noncompliance would be a measurement of placebo compliance during a run-in period of the study. Indeed, run-in placebo compliance is associated with treatment compliance and likely not further related to the treatment effect, given the actual compliance during the active study period. More generally, one can use a second causal IV as an IV for the measurement error. Indeed, an IV-C is associated with the considered exposure by Assumption A1 and does not modify the target causal effect by Assumption A3. It thus satisfies the conditions for an IV-M. The use of multiple IV-C's turns out feasible in practice as it is commonly considered in econometrics and more recently also in Mendelian randomization studies (Didelez and Sheehan, 2007). For instance, to assess the effect of C-reactive protein on insulin resistance, one may use the CRP-gene as an IV-C and the interleukin-6 gene - which is known to be associated with C-reactive protein through other pathways than the CRP-gene and which thus applies as a second IV-C - as an IV-M. Note, however that the restrictions for an IV-M are much weaker than those for an IV-C as an IV-M need not satisfy the exclusion restriction, nor the randomization assumption (see Assumption A1). Note also that assumption (4.4) is weaker than the typical IV-assumption encountered in measurement error models (Carroll et al., 2006) as it does allow for the IV to be associated with the outcome, conditional on the exposure.

*Assumption A5 (Constant average measurement error):* For simplicity and because information about the average error is weak, we develop our approach below for constant (but unknown) average error $E(W_i - Z_i|\mathbf{X}_i, R_i) = \delta^*$. This assumption is standard in the measurement error literature, but is straightforwardly relaxed (e.g., by postulating $E(W_i - Z_i|\mathbf{X}_i, R_i) = \delta_0^* + \delta_1^{*'}\mathbf{X}_i$).

## 2.2 Inference

Our goal is to estimate the parameter $\boldsymbol{\psi}^*$ indexing (4.2) under model $\mathcal{A}$, which is the model for the observed data $(Y_i, W_i, R_i, \mathbf{X}_i)$ defined by assumptions A1-A5 with the conditional density

$$f(R_i|\mathbf{X}_i) \text{ known.} \tag{4.5}$$

The latter assumption holds in a randomized trial when $R_i$ indicates randomized assignment, because treatment allocation is then under the control of the investigator. If assumption (4.5) fails, then all further results continue to hold upon replacing $f(R_i|\mathbf{X}_i)$ with a consistent estimator.

It follows from Proposition 1 below that the average measurement error $\delta^*$ is all that must be known for identifying $\boldsymbol{\psi}^*$ under model $\mathcal{A}$.

*Proposition 1.* Model $\mathcal{A}$ is the same model for the observed data as the conditional mean independence model $\mathcal{B}$ for the observed data model, which is defined by a known function $f(R_i|\mathbf{X}_i)$ (as in (4.5)) and

$$E\left[Y_i - \gamma(\mathbf{S}_i;\boldsymbol{\psi}^*)\left(W_i - \delta^*\right)|\mathbf{X}_i, R_i\right] = E\left[Y_i - \gamma(\mathbf{S}_i;\boldsymbol{\psi}^*)\left(W_i - \delta^*\right)|\mathbf{X}_i\right]. \tag{4.6}$$

Note the essential difference between models $\mathcal{A}$ and $\mathcal{B}$. Model $\mathcal{A}$ is expressed in terms of counterfactuals and therefore has parameters with a causal interpretation. Model $\mathcal{B}$ imposes the same restrictions on the observed data as model $\mathcal{A}$, but is not expressed in terms of counterfactuals. This makes the parameters in this model harder to interpret, but simplifies inference as the model is expressed in terms of observed data only. Note also that model $\mathcal{A}$ imposes only weak restrictions on the error distribution. First, it allows the error to be associated with both the true exposure $Z_i$ and observed exposure $W_i$. It thus encompasses both the classical and Berkson error model (Carroll et al., 2006). In addition, by avoiding assumptions about the conditional association between $W_i$ and $Y_i$, given $Z_i$, it allows for so called differential error, which is associated with outcome conditional on exposure (see the proof of Proposition 1 for a more explicit argument). This can be important. For instance, in a clinical trial, patients may be more reluctant to 'confess' to noncompliance when their outcome stayed below target. Finally, model $\mathcal{A}$ makes no assumptions on the measurement error distribution other than Assumption 5. This is useful because the error distribution can be quite complex. For instance, with low level exposures negative errors become constrained by the fact that negative exposures are never reported.

By Proposition 1 and the fact that $\boldsymbol{\psi}^*$ is the same functional of the observed data under models $\mathcal{A}$ and $\mathcal{B}$, inference for $\boldsymbol{\psi}^*$ is the same under both models. It follows that the set of all consistent and asymptotically normal (CAN) estimators for $\boldsymbol{\psi}^*$ is the same under models $\mathcal{A}$ and $\mathcal{B}$, where the latter can be obtained as in Robins (1994) by solving the mean independence estimating equations

$$\sum_{i=1}^{n} \mathbf{d}(R_i, \mathbf{X}_i)\left[Y_i - \gamma(\mathbf{S}_i;\boldsymbol{\psi})(W_i - \delta) - q(\mathbf{X}_i)\right] = 0 \tag{4.7}$$

jointly for $\boldsymbol{\theta} = (\boldsymbol{\psi}', \delta)'$, with $\mathbf{d}(R_i, \mathbf{X}_i) = \mathbf{g}(R_i, \mathbf{X}_i) - \mathrm{E}\left\{\mathbf{g}(R_i, \mathbf{X}_i)|\mathbf{X}_i\right\}$ and with $\mathbf{g}(R_i, \mathbf{X}_i)$ and $q(\mathbf{X}_i)$ arbitrary (non-trivial) index functions of the dimension of $\boldsymbol{\theta}$.

Note that estimating equation (4.7) is designed to make the predicted exposure-free outcomes $Y_i - \gamma(\mathbf{S}_i;\boldsymbol{\psi})(W_i - \delta)$ mean independent of $R_i$, conditional on $\mathbf{X}_i$, in order to satisfy Assumption A1. The index functions $\mathbf{g}(R_i, \mathbf{X}_i)$ and $q(\mathbf{X}_i)$ can be arbitrarily chosen without affecting the consistency of the resulting estimators of $\boldsymbol{\psi}^*$. In particular, they can be chosen in view of efficiency. Under the homoscedasticity assumption that the conditional variance of $Y_i - \gamma(\mathbf{X}_i;\boldsymbol{\psi})(W_i - \delta)$, given $(R_i, \mathbf{X}_i)$, is constant, semi-parametric efficiency (Robins, 1994) is for instance obtained by setting $q(\mathbf{X}_i)$ equal to

$$q_{opt}(\mathbf{X}_i) = E\left\{Y_i - \gamma(\mathbf{S}_i;\boldsymbol{\psi})(W_i - \delta)|\mathbf{X}_i, R_i\right\}$$

and $\mathbf{d}(R_i, \mathbf{X}_i)$ equal to $\mathbf{d}_{opt}(R_i, \mathbf{X}_i) = \mathbf{g}_{opt}(R_i, \mathbf{X}_i) - \mathrm{E}\left\{\mathbf{g}_{opt}(R_i, \mathbf{X}_i)|\mathbf{X}_i\right\}$ with

$$\mathbf{g}_{opt}(R_i, \mathbf{X}_i) \quad = \quad E\left\{\frac{\partial\gamma(\mathbf{S}_i;\boldsymbol{\psi})(W_i - \delta)}{\partial\boldsymbol{\theta}}|\mathbf{X}_i, R_i\right\}.$$

These choices will be used later in the data analysis and simulation study.

*Theorem 1.*

1. Under weak regularity conditions, the solution $\hat{\boldsymbol{\psi}}(\mathbf{d}, q)$ to (4.7) satisfies $\sqrt{n}\left(\hat{\boldsymbol{\psi}}(\mathbf{d}, q) - \boldsymbol{\psi}^*\right) \rightarrow N\left(0, \Gamma(\mathbf{d}, q)\right)$ in distribution, where

$$\Gamma(\mathbf{d}, q) = E^{-1}\left\{\frac{\partial\mathbf{U}_i(\mathbf{d}, q;\boldsymbol{\psi}^*)}{\partial\boldsymbol{\psi}}\right\} Var\{\mathbf{U}_i(\mathbf{d}, q;\boldsymbol{\psi}^*)\} E^{-1\prime}\left\{\frac{\partial\mathbf{U}_i(\mathbf{d}, q;\boldsymbol{\psi}^*)}{\partial\boldsymbol{\psi}}\right\} \tag{4.8}$$

with $\mathbf{d}(R_i, \mathbf{X}_i) = (\mathbf{d}_\psi(R_i, \mathbf{X}_i), d_\delta(R_i, \mathbf{X}_i))$ and

$$\begin{aligned}\mathbf{U}_i(\mathbf{d}, q;\boldsymbol{\psi}) \quad = \quad & \left[\mathbf{d}_\psi(R_i, \mathbf{X}_i) - \frac{E\left\{\mathbf{d}_\psi(R_i, \mathbf{X}_i)\gamma(\mathbf{S}_i;\boldsymbol{\psi})\right\}}{E\left\{d_\delta(R_i, \mathbf{X}_i)\gamma(\mathbf{S}_i;\boldsymbol{\psi})\right\}} d_\delta(R_i, \mathbf{X}_i)\right] \\ & \times \left[Y_i - \gamma(\mathbf{S}_i;\boldsymbol{\psi})(W_i - \delta) - q(\mathbf{X}_i)\right]\end{aligned}$$

2. The average error $\delta^*$ is not root-$n$ estimable at $\boldsymbol{\psi}^* = 0$.

3. For arbitrary $(\mathbf{d}, q)$, $\Gamma(\mathbf{d}_{opt}, q_{opt}) \leq \Gamma(\mathbf{d}, q)$ where $A \leq B$ is defined as $A - B$ being semi-positive definite.

Part 1 of Theorem 1 confirms that the solution $\hat{\boldsymbol{\psi}}(\mathbf{d}, q)$ to (4.7) is a root-$n$ CAN estimator of $\boldsymbol{\psi}^*$. This is even so at $\boldsymbol{\psi}^* = 0$ where $\delta^*$ is not root-$n$ estimable. Theorem 1 also shows how to calculate the efficient score $\mathbf{U}_i(\mathbf{d}_{opt}, q_{opt};\boldsymbol{\psi})$ for $\boldsymbol{\psi}^*$ in model $\mathcal{A}$. For example, in Section 4, we will consider the analysis of a placebo-controlled randomized trial with $Z_i$ denoting compliance to the experimental treatment. Because the placebo arm ($R_i = 0$) is unexposed, $Z_i = Z_i R_i$ and there is no measurement error in that arm so that we modify Assumption A5 to $E(W_i - Z_i|\mathbf{X}_i, R_i) = \delta^* R_i$. With $\mathbf{X}_i = \mathbf{T}_i$, $\gamma(\mathbf{S}_i;\psi) = \psi$ and assuming homoscedasticity and constant randomization

probabilities $\pi = P(R_i = 1) = P(R_i = 1|\mathbf{X}_i)$, the semi-parametric efficient score for $\psi^*$ is

$$(R_i - \pi)\left[E(W_i|R_i = 1, \mathbf{X}_i) - E\left\{E(W_i|R_i = 1, \mathbf{X}_i)\right\}\right]\left\{Y_i - \psi(W_i - \delta)R_i - q_{opt}(\mathbf{X}_i)\right\}.$$

This score differs from the efficient score in the absence of biased measurement error (i.e. assuming that $\delta^* = 0$) in that it carries the additional term $E\left\{E(W_i|R_i = 1, \mathbf{X}_i)\right\}$, which corrects for estimation of the error mean. This term reduces the variance of the estimating functions and, as such, encodes efficiency loss. Specifically, note that the efficient score becomes 0 when the IV-M, $T$, is uncorrelated with the observed exposure, and hence that $\psi^*$ is not root-$n$ estimable in that case. By the same token, instruments for the measurement error which are weakly correlated with observed exposure, may yield unstable effect estimates.

## 2.3 Bias-variance Trade-off

The anticipated loss of efficiency of the error-adjusted estimator raises the question whether the bias correction developed so far is useful. To this end, we investigate the bias-variance trade-off for the error-adjusted and the standard unadjusted estimator for the causal effect $\psi^*$, in a specific case. Tractable expressions for the mean-squared error of both estimators, are obtained when $Z \sim N(\mu_z, \sigma_z^2), T|Z \sim N(\nu_0 + \nu_1 Z, \sigma_{t|z}^2), Y_0|Z, T \sim N(\alpha_0 + \alpha_1 Z + \alpha_2 T, \sigma_0^2)$ and $Y = Y_0 + (\psi + \epsilon)RZ$ with $\epsilon|Y_0, Z, T \sim N(0, \sigma^2)$. Under the working assumption of no systematic measurement error (i.e. fixing $\delta^* = 0$ in equation (4.7) and not estimating it), the efficient score for $\psi^*$ is $U_u(\psi) = (0.5 - R)\mathrm{E}(W|T, R = 1)\{Y - \psi RW - E(Y|R = 0, T)\}$ in model $\mathcal{A}$ with $\mathbf{X}_i = \mathbf{T}_i$ under the above data-generating mechanism. It follows after some algebra that the solution $\hat{\psi}_u$ to $\sum_{i=1}^n U_{ui}(\psi) = 0$ has bias which can be approximated by

$$\mathrm{E}^{-1}\left(\frac{\partial U_u(\psi)}{\partial \psi}\right)\mathrm{E}\{U_u(\psi)\} = \frac{\psi\delta(\mu_z + \delta)}{\sigma_z^2 - \sigma_{z|t}^2 + (\mu_z + \delta)^2}$$

where $\sigma_{z|t}^2 = \sigma_z^2\sigma_{t|z}^2/(\nu_1^2\sigma_z^2 + \sigma_{t|z}^2)$ is the conditional variance of $Z$ given $T$, and asymptotic variance given by

$$\frac{1}{n}\left[\frac{4\sigma_0^2 + 4\alpha_1^2\sigma_{z|t}^2 + 2\psi^2\sigma_u^2 + \psi^2\delta^2}{\sigma_z^2 - \sigma_{z|t}^2 + (\mu_z + \delta)^2} + \frac{\psi^2\delta^2(\sigma_z^2 - \sigma_{z|t}^2)}{\left\{\sigma_z^2 - \sigma_{z|t}^2 + (\mu_z + \delta)^2\right\}^2}\right].$$

Allowing for systematic measurement error, the efficient estimator $\hat{\psi}_c$ for $\psi^*$ under model $\mathcal{A}$ has no asymptotic bias and asymptotic variance which equals

$$\frac{1}{n}\frac{\sigma_0^2 + \alpha_1^2\sigma_{z|t}^2 + 0.5\psi^2\sigma_u^2}{0.5^2(\sigma_z^2 - \sigma_{z|t}^2)}.$$

Note that the bias of the unadjusted estimator and the asymptotic variance of both estimators is inversely proportional to the multiple correlation coefficient for the regression of $Z$ on $T$. The variance of the error-adjusted estimator becomes infinite when $Z$ and $T$ are uncorrelated.

Figure 1 shows the range of values $\delta$ for the average error under which the standard estimator, which ignores measurement error, has smaller mean squared error than the error-adjusted estimator. This is displayed in function of the sample size and the correlation between $Z$ and $T$. Specifically, the values of $\delta$ comprised between the solid lines indicate data-generating mechanisms under which the standard estimator outperforms the error-adjusted estimator in terms of mean squared error. The figure was constructed using parameter values which are reflective of the hypertension study that we will analyze in Section 4. It shows that at small sample sizes ($n = 105$), correction for systematic measurement error leads to smaller mean squared error, but only when the systematic error component is substantial (i.e. of about the size of the average exposure $\mu_z$) and, at the same time, the IV-M, $T$, is strongly correlated with $Z$. Figure 1 reveals further that bias correction using the error-adjusted estimator is only helpful at moderate degrees of error and moderate correlations between $T$ and $Z$ when sample sizes are very large.

# 3 Incorporating prior information

The previous results demonstrate the poor performance of the error-adjusted estimator, even in settings where the sample size is moderate and good (pre-exposure) predictors of exposure are available. In particular, tests of the causal null hypothesis can loose substantial power by using this approach rather than the standard test of the causal null (i.e. that $R$ and $Y$ are conditionally independent, given $\mathbf{X}_i$), which is immune to measurement error on the exposure (Goetghebeur and Vansteelandt, 2005). This is surprising, considering that the score test of $\boldsymbol{\psi}^* = \mathbf{0}$ under model $\mathcal{A}$ does not involve $\delta^*$ and, hence, does not need to correct for measurement error when testing the causal null hypothesis. Curiously, it follows that one can validly and efficiently test the causal null hypothesis without correction for measurement error, but that a score test of $\boldsymbol{\psi}^* = \boldsymbol{\psi}_0$ with $\boldsymbol{\psi}_0$ arbitrarily close to (but different from) 0,

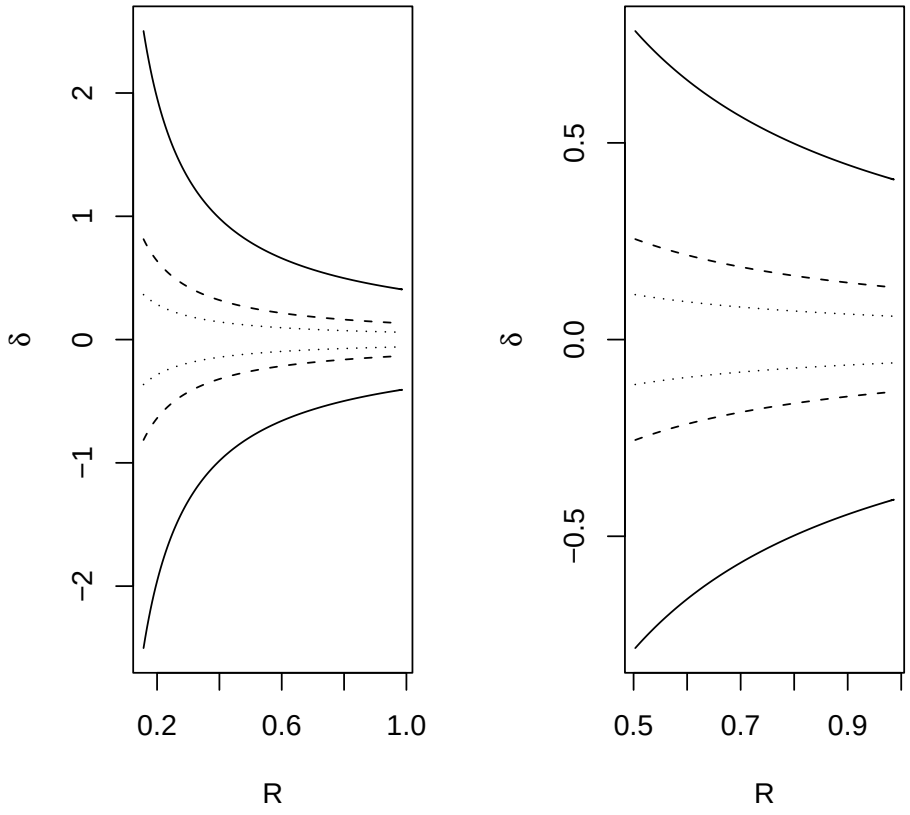



Figure 1: *Curves indicating the tuples* $(R, \delta)$ *where the standard SMM estimator and the error-adjusted instrumental variable estimator have the same mean squared error, for* $R$ *equalling the correlation between* $Z$ *and* $T$*, for different sample sizes* $n = 105, 1000$ *and 5000 and with* $\mu_z = 0.85, \sigma_z^2 = 0.11, \nu_0 = 0.75, \nu_1 = 0.12, \sigma_{t|z}^2 = 0.012, \alpha_0 = -4.4, \alpha_1 = 6.8, \alpha_2 = -13.7, \sigma_0^2 = 53.2, \sigma_u^2 = 0, \psi = -7.5$ *and* $\sigma^2 = 0$*. Left: for* $R$ *from 0 to 1; Right: for* $R$ *from 0.5 to 1.*

would require correcting for measurement error and hence could imply a serious and sudden loss of power.

The root cause of this apparent discontinuity is the fact that, as shown in Part 2 of Theorem 1, $\delta^*$ is not root-$n$ estimable at $\boldsymbol{\psi}^* = \mathbf{0}$ so that estimation of $\delta^*$ affects the distribution of the score test statistic, even though it gets multiplied by $\boldsymbol{\psi}^* = \mathbf{0}$ in the test statistic (i.e. even at the causal null hypothesis). In particular, it follows from the proof of Theorem 1 that $\sqrt{n}\left\{\hat{\delta}(\mathbf{d}, q) - \delta^*\right\}\boldsymbol{\psi}^*$, with $\hat{\delta}(\mathbf{d}, q)$ the solution for $\delta$ to (4.7), is bounded in probability with strictly positive variance for each value of $\boldsymbol{\psi}^*$, suggesting that $\hat{\delta}\boldsymbol{\psi}^*$ fluctuates around 0, even when $\boldsymbol{\psi}^* = \mathbf{0}$. This happens with decreasing variance as the sample size increases.

Similar problems of inestimability at a local point in the parameter space have been noted in other measurement error problems (Gustafson, 2005). More general problems of inferring a parameter $\boldsymbol{\psi}^*$ when a nuisance parameter $\delta^*$ disappears under the null ($\boldsymbol{\psi}^* = \mathbf{0}$) have been discussed mainly in the econometrics literature (Davies, 1977, 1987; Hansen, 1992; Andrews and Ploberger, 1994). To the best of our knowledge, attention has only been given to testing problems in which the test statistic involves a nuisance parameter which is unidentified at the null. Some of these approaches assume that the nuisance parameter lies within a known open set and base inference on the supremum of a score or likelihood ratio test statistic, taken over all values of the nuisance parameters in the chosen set (Davies, 1977, 1987). Andrews and Ploberger (1994) postulate a prior distribution for the nuisance parameter and base inference on the average of a score or likelihood ratio test statistic over the chosen prior distribution. Our problem is different in that our main focus is on estimation rather than testing, and that a score test for the causal null hypothesis does not involve the nuisance parameter. Nonetheless, inspired by the work of Davies (1977, 1987) and by sensitivity analyses for IV-estimators with measurement error (Goetghebeur and Vansteelandt, 2005), we proceed by considering estimation under the assumption that the average error $\delta^*$ lies within a known open set $\Delta$. This strategy is further motivated by the fact that (a) subject-matter experts often have a good sense of the extent of expected mismeasurement (Gustafson, 2005); (b) it forces the estimate for $\delta^*$ to have bounded variation around the truth, contrary to what happens under the approach of Section 2.2.

## 3.1 Improved Error Adjustment

Our first approach under the assumption that $\delta^* \in \Delta = ]\Delta_l, \Delta_u[$ solves equations (4.7) with $\delta$ replaced by $\{I(\lambda < 0)\Delta_l + I(\lambda > 0)\Delta_u\}\lambda/(1 + |\lambda|)$ and $\lambda$ unknown. This guarantees estimates $\delta^*(\hat{\lambda})$ within the set $\Delta$ and thus greatly improves the stability

of estimators for the causal effect $\boldsymbol{\psi}^*$. A drawback which will become apparent in the simulation study of Section 3.2, is that tests of the causal null hypothesis may still loose substantial power under this approach due to the fact that also $\lambda$ is not root-$n$ estimable at $\boldsymbol{\psi}^* = \mathbf{0}$. To accommodate this, we will develop a second, recommended approach which trades bias for precision by solving a weighted average of the estimating functions for the standard SMM estimator and for the error-adjusted estimator of Section 2.3. Estimating functions for the standard estimator are weighted proportionally to the estimated probability that $\delta^*$ falls outside the chosen set $\Delta$. The philosophy behind this choice is that estimates for $\delta^*$ will not likely fall within the set $\Delta$ in situations where little information on the error mean is available. Hence more weight will be given to the standard unadjusted estimator in those cases. For pedagogic purposes, we will explain our proposal for the case $\gamma(\mathbf{X}_i;\psi) = \psi$ and with Assumption A5 modified to $E(W_i - Z_i|\mathbf{X}_i, R_i) = \delta^* R_i$. We further delete reference to the index functions $(d, q)$ in the estimators. For each value $\psi$ in a chosen grid, we calculate an estimator $\hat{\delta}(\psi)$ for $\delta^*$ which solves (4.7) for the given $\psi$ with $d_\delta(R_i, \mathbf{X}_i)$ in place of $d(R_i, \mathbf{X}_i)$. Next, we consider a weighted average of the estimating function $U_{\psi i}(\psi, \delta)$ for $\psi^*$ (as defined in (4.7) with $d_\psi(R_i, \mathbf{X}_i)$ in place of $d(R_i, \mathbf{X}_i)$), evaluated at the profile estimator $\delta = \hat{\delta}(\psi)$ and at $\delta = 0$, respectively:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{U}_i(\psi) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{P}\{\hat{\delta}(\psi) \in \Delta\} U_{\psi i}\{\psi, \hat{\delta}(\psi)\} + \hat{P}\{\hat{\delta}(\psi) \notin \Delta\} U_{\psi i}(\psi, 0) \quad (4.9)$$

In this expression, the weights involve the estimated probability $\hat{P}\{\hat{\delta}(\psi) \notin \Delta\}$ that $\hat{\delta}(\psi)$ falls outside the chosen interval $\Delta = ]\Delta_l, \Delta_u[$. Using a similar development as in the proof of Theorem 1, this probability can be approximated by

$$P\{\hat{\delta}(\psi) \notin \Delta\} = 1 + \Phi\left(\frac{\Delta_l - \delta}{\sigma(\psi)/(\sqrt{n}|\psi|)}\right) - \Phi\left(\frac{\Delta_u - \delta}{\sigma(\psi)/(\sqrt{n}|\psi|)}\right) \quad (4.10)$$

with $\Phi(.)$ the cumulative standard normal distribution function, and where $\delta$ may be replaced by a consistent estimator (for instance, $\min[\Delta_u, \max\{\Delta_l, \hat{\delta}(\psi)\}]$) and $\sigma(\psi)$ by a consistent estimator $\hat{\sigma}(\psi)$ for the standard deviation of the scaled estimating function $E^{-1}\left[d_\delta(R, T, X)R\right] U_{i\delta}(\psi, \delta)$ for $\delta^*$. We define the improved error-adjusted estimator $\tilde{\psi}$ for $\psi^*$ as the value of $\psi$ at which the score test (4.9) becomes zero. Curiously, this estimator assigns much weight to the standard estimating equations (which do not adjust for measurement error) when the error mean is estimated to be large. This is (a) because the philosophy behind the estimator is that such large values for the error mean are indicative of imprecision; and (b) because the estimating functions are designed to equal the unadjusted estimating functions at the causal null hypothesis (see further).

*Theorem 2.* Suppose that $\gamma(\mathbf{X}_i;\psi) = \psi, Z_i = Z_i R_i$ and $E(W_i - Z_i|\mathbf{X}_i, R_i) = \delta^* R_i$. Then, under regularity conditions stated in the Appendix and for any fixed $\psi$, $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{U}_i(\psi) \rightarrow N(0, \Sigma(\psi))$ in distribution, where $\Sigma(\psi)$ is the variance of

$$P\{\hat{\delta}(\psi) \in \Delta\}U_{i\psi}(\psi,\delta) + P\{\hat{\delta}(\psi) \notin \Delta\}U_{i\psi}(\psi,0) - \Big[P\{\hat{\delta}(\psi) \in \Delta\}$$
$$+ \left\{\varphi\left(\frac{\Delta_l - \delta}{\sigma(\psi)/(\sqrt{n}|\psi|)}\right) - \varphi\left(\frac{\Delta_u - \delta}{\sigma(\psi)/(\sqrt{n}|\psi|)}\right)\right\}\frac{\sqrt{n}|\psi|\delta}{\sigma(\psi)}\Big]$$
$$\times \frac{E\{d_\psi(R,\mathbf{X})R\}}{E\{d_\delta(R,\mathbf{X})R\}}U_{i\delta}(\psi,\delta)$$

with $\varphi(.)$ the standard normal density function.

Theorem 2 can be used to construct $(1-\alpha)100\%$ confidence intervals for $\psi^*$ as the range of values $\psi_0$ for $\psi$ such that the two-sided score test based on (4.9) does not reject the null hypothesis $H_0 : \psi^* = \psi_0$ at the $\alpha 100\%$ significance level. To evaluate this score test, one may replace the variance of the score test statistic by the sample variance with $P\{\hat{\delta}(\psi) \in \Delta\}$ replaced by $\hat{P}\{\hat{\delta}(\psi) \in \Delta\}, \delta$ by $\hat{\delta}(\psi)$ and $\sigma(\psi)$ by $\hat{\sigma}(\psi)$. The resulting confidence intervals have the desirable feature that, asymptotically, they exclude 0 if and only if the standard test of the causal null hypothesis (i.e., that $Y \perp\!\!\!\perp R|\mathbf{X}$) rejects. Indeed, at the null hypothesis $\hat{P}\{\hat{\delta}(0) \notin \Delta\} \xrightarrow{p} 1$ and hence the score test statistic becomes

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}U_{\psi i}(\psi,0) \quad = \quad \frac{1}{\sqrt{n}}\sum_{i=1}^{n}d_\psi(R_i,\mathbf{X}_i)\{Y_i - q(\mathbf{X}_i)\} + o_p(1)$$

for an arbitrary function $d_\psi(R_i,\mathbf{X}_i)$ with conditional mean zero, given $\mathbf{X}_i$. When $\hat{\delta}(\psi)$ in $U_{\psi i}(\psi,\hat{\delta}(\psi))$ is restricted to $\Delta$ as described in the first paragraph of this Section, this statistic is asymptotically equivalent to a score test statistic of the causal null hypothesis under the observed data model defined by restriction (4.5).

Unfortunately, the suggested confidence intervals are no uniform asymptotic confidence intervals. The reason is that, at each sample size, there exists a $\psi^*$ depending on $n$ which is sufficiently close to zero that the score test statistic (4.9) is significantly biased as a result of bias in the estimating functions of the standard unadjusted SMM estimator. Specifically, it follows from the proof of Theorem 2 that the improved error-adjusted estimator $\tilde{\psi}$ is asymptotically biased within root-$n$ shrinking neighbourhoods of zero (i.e. when $\psi^* = k/\sqrt{n}$ for some constant $k$) and may not converge to a normal distribution along such sequences. Curiously, $\tilde{\psi}$ is asymptotically unbiased and normally distributed along faster converging sequences (i.e. when $\psi^* = kn^{-a}$ for some constant $k$ and $a > 1/2$) and in particular at $\psi^* = 0$. The reason is that, although

the probability that $\hat{\delta}(\psi) \in \Delta$ now converges to 0 and hence $\tilde{\psi}$ is asymptotically equivalent to the standard unadjusted SMM estimator, $\psi^*$ is sufficiently close to zero to make any bias in the estimator negligible. Likewise, $\tilde{\psi}$ is asymptotically unbiased and normally distributed along slower converging sequences (i.e. when $\psi^* = kn^{-a}$ for some constant $k$ and $0 \leq a < 1/2$). The reason is that the estimated probability of $\hat{\delta}(\psi) \in \Delta$ now converges to 1 so that the improved error-adjusted estimator is asymptotically equivalent to the error-adjusted estimator of Section 2.2, which is asymptotically unbiased.

The practical implication of the foregoing discussion is that the improved error-adjusted estimator $\tilde{\psi}$ and confidence intervals have no guaranteed performance in finite samples in the sense that, for each sample size, one can find a causal effect $\psi^*$ which is close, but not too close to zero so that $\tilde{\psi}$ is significantly biased and that confidence intervals for $\psi^*$ do not cover $\psi^*$ at the nominal level. This local bias is the price we pay for estimators with smaller variability and limited loss of power for testing the causal null hypothesis. Because this problem only appears within 1 over root-$n$ distances from zero and not within larger or shorter distances, we expect adequate performance in many practical situations. However, to be conservative we develop uniform asymptotic confidence intervals in the next section.

## 3.2 Uniform Asymptotic Confidence Intervals

Uniform asymptotic $(1-\alpha)100\%$ confidence intervals are expected to have better finite-sample properties than the intervals of the previous section because they guarantee the existence of a minimal sample size such that, at larger sample sizes, they cover $\psi^*$ with at least $(1-\alpha)100\%$ chance regardless of the value of $\psi^*$. Following ideas in Robins (2005), we construct such intervals by first constructing, for each $\psi$, an asymptotic uniform $(1-\epsilon)100\%$ confidence interval $C(\psi)$ for $\delta^*$, where the choice of $\epsilon < \alpha$ will be discussed later. Because we assume the parameter space for $\delta^*$ to be $\Delta$, a conservative asymptotic interval $C(\psi)$ may be obtained as

$$\left\{\hat{\delta}(\psi) \pm z_{\epsilon/2}\frac{\hat{\sigma}(\psi)}{|\psi|\sqrt{n}}\right\} \cap \Delta.$$

Using Theorem 5.1 in Robins (2005), an asymptotic uniform $(1-\alpha)100\%$ confidence interval for $\psi^*$ may be obtained as the set of $\psi$-values for which

$$\inf_{\delta \in C(\psi)} |Var^{-1/2}\{U_{\psi i}(\psi,\delta)\}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_{\psi i}(\psi,\delta)| < z_{(\alpha-\epsilon)/2}$$

The optimal choice of $\epsilon$ that leads to confidence intervals of minimum length is difficult to determine (Robins, 2005). In this article, we propose to choose $\epsilon$ in function of

$\psi$ as $0.5\alpha|\psi|/(1+|\psi|)$. This choice guarantees that $C(\psi)$ will equal $\Delta$ for $\psi^* = 0$ and equal a $(1-\alpha/2)100\%$ confidence interval for $\delta^*$ at causal effects $\psi^*$ far from 0. The philosophy behind this choice is that estimates for $\delta^*$ will be highly imprecise at causal effects close to zero and hence, given that the parameter space for $\delta^*$ is bounded, we expect no difference between 100% confidence intervals and $(1-\alpha)100\%$ confidence intervals for $\delta^*$ at $\psi^* = 0$. As such, we need not offer the significance level for $\psi^*$ at small causal effects and will thus get narrower intervals in return. Specifically, the proposed confidence intervals have the feature that they involve no correction for measurement error at $\psi^* = 0$, which is desirable because there is no bias due to measurement error at $\psi^* = 0$.

# 4 Data analysis

We analyze data from a placebo-controlled randomized hypertension trial which enrolled some 300 hypertensive patients (Goetghebeur and Lapp, 1997). After a run-in period of 4 weeks where all patients received placebo tablets, they were randomized to 4 weeks of one of two active treatments (A or B) or placebo. All treatments were prescribed at one tablet per day. Here, we analyze the subset of 105 patients randomized to A or placebo, for whom treatment compliance was electronically measured, ignoring 5 patients with missing diastolic blood pressure or compliance.

An intent-to-treat analysis reveals an average difference in blood pressure reduction over the active 4-week study period of 7.5 mmHg (95% CI 4.0; 11.0) without adjustment. This reveals the effect of assignment to treatment A (instead of placebo) on expected diastolic blood pressure reduction from baseline (i.e. the time of randomization). Primary interest lies however in the effect of *received* treatment on average blood pressure reduction. We will therefore fit model (4.3) with $Y_i$ the blood pressure reduction over the active study period, $R_i$ the randomization indicator as the IV-C (which is 1 if assigned to experimental treatment and 0 if assigned placebo), $Z_i$ the average number of prescribed pills taken, and $\mathbf{X}_i$ the age of patient $i$. Assuming that compliance measurements are free of systematic error, we estimate that the average blood pressure reduction would have been 9.6 mmHg (95% CI 3.5; 11.8) smaller over the study period among those who choose to take on average one pill per day, had they not taken the exposure. Note that this estimand averages the effect over patients with different compliance patterns, but with the same average pill intake. Distinguishing between these patients would require more detailed compliance measures, but would suffer from identifiability problems (Vansteelandt and Goetghebeur, 2005).

In reality, there are concerns that also electronic compliance measurements carry systematic errors and thus that the above estimate may be biased. Because this study

was not designed to correct for measurement error, no natural instrumental variables for the measurement error have been recorded. Our analysis is hence for illustrative purposes only and will use age as an IV-M (i.e., $\mathbf{T}_i = \mathbf{X}_i$ equals age). Age was chosen because effect modification through age is not anticipated (nor observed) in this study population, which consists of middle aged hypertensive patients (5th, 95th percentiles: 41 and 69 years), and thus Assumption A4 is anticipated to be approximately true. A more adequate analysis would use placebo compliance during the run-in period (where no electronic adherence measures were taken) as an IV-M. Using the error-adjusted estimator of Section 2.2, we estimate a larger treatment effect of 27.0 mmHg (95% CI -91.2; 145.2). To improve this imprecise result, we impose the weak assumption that the average error is smaller than 0.25. We believe this assumption to be reasonable, given that the observed percentage of assigned dose taken (i.e. the observed exposure) is 0.85 (i.e., 85%) on average. Choosing $\Delta = [-0.25, 0.25]$ thus allows for 30% of the observed average exposure to be due to systematic error. Using the improved error-adjusted estimator for inference, we estimate a slightly smaller effect of 9.0 mmHg (95% CI 4.4; 17.4) as compared to the standard analysis. As predicted by the theory, the estimate is less precise than the unadjusted estimator, but still significantly different from 0 at the 5% significance level. The uniform asymptotic 95% confidence interval (2.7; 16.8) has a more guaranteed performance in finite samples. To investigate the sensitivity of our result to the choice of $\Delta$, Figure 2 shows the improved error-adjusted estimate, along with uniform 95% confidence intervals in function of the assumed maximum error mean $\Delta_u$, with $\Delta = [-\Delta_u, \Delta_u]$. It reveals reasonable stability. Comparison with the sensitivity analysis results of Goetghebeur and Vansteelandt (2005) shows that the error-adjustment described in this article reduces uncertainty.

# 5 Simulation study

To investigate the behaviour of the error-adjusted estimators in finite samples with $\psi^*$ possibly close to zero, we conducted simulation experiments. Each experiment was based on 5000 replications of random samples of size 105 (i.e. the sample size of the blood pressure study) or 1000, generated as follows. In each experiment, the instrument $T$ for the measurement error was normal with mean 0.83 and standard deviation 0.14 and $R$ was independently generated from a Bernoulli distribution with success probability 0.5. The true exposure $Z$ and exposure-free response were generated as $Z = T + 0.32\epsilon_Z$ and $Y_0 = -4.4 + 6.8Z - 7.3T + 7.3\epsilon_0$ for independent standard normal variates $\epsilon_Z, \epsilon_0$. Finally, we generated $Y$ as $Y_0 + \psi RZ$ and the observed exposure $W$ as $W = Z + U$ where $U \sim N(\delta, 0.01)$. Table 4.1 summarizes the

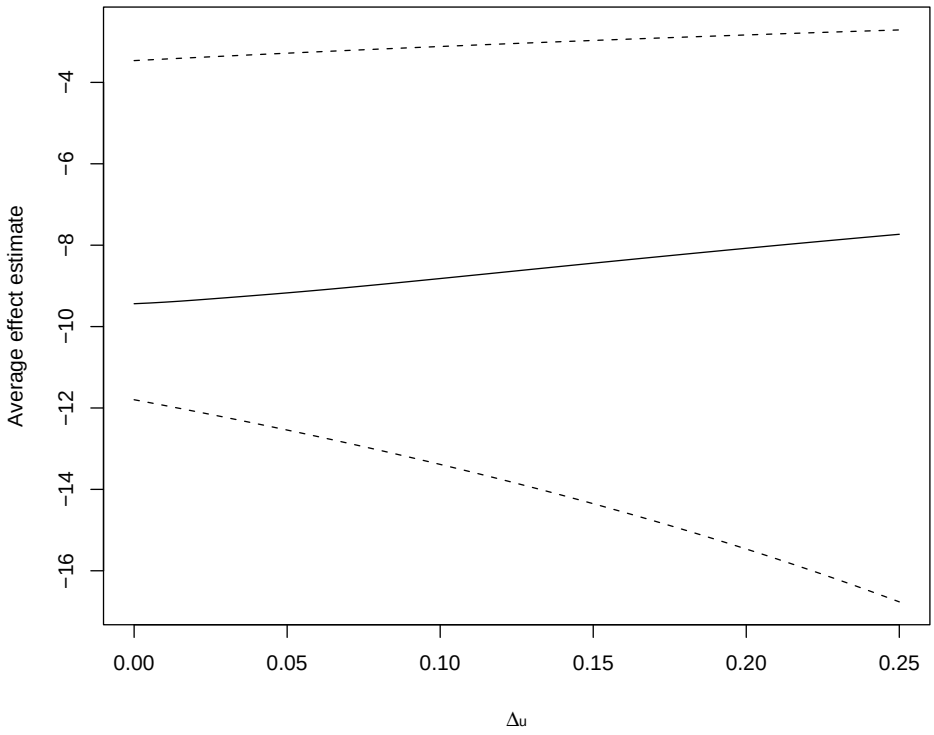



Figure 2: *Improved error-adjusted estimate, along with uniform 95% confidence intervals in function of the maximum error mean* $\Delta_u$*, with* $\Delta = [-\Delta_u, \Delta_u]$.

results for estimation of $\psi$ using i) the standard IV estimator which ignores systematic measurement error (STD); (ii) the error-adjusted estimator of Section 3.1 (IV1); iii) the error-adjusted estimator of Section 3.3 which guarantees estimates for $\delta$ to stay within $\Delta = [\Delta_l, \Delta_u]$ with $\Delta_u = -\Delta_l$ equal to 0.5, 0.25 or 0.05, by defining $\delta = \{I(\lambda < 0)\Delta_l + I(\lambda > 0)\Delta_u\}\lambda/(1 + |\lambda|)$ for unknown $\lambda$ (IV2); the improved error-adjusted estimator of Section 3.3 with the same choices for $\Delta$ (IV3). In addition, the table shows uniform asymptotic 95% confidence intervals (UI) corresponding to these choices. The results for the different estimators are as predicted by the theory. The error-adjusted estimator (IV1) is extremely variable at small sample sizes, but performs adequately at larger sample sizes, even at $\psi = 0$. Estimator (IV2) is less variable, although still substantially less precise than the standard unadjusted estimator. Figures 3 and 4 show that estimator (IV1) is normally distributed in moderate sample sizes, even at $\psi = 0$, but not in small samples. It also shows that the improved error-adjusted estimator (IV3) is much less variable than the error-adjusted estimator (IV1). While the former follows a normal distribution in small samples, deviations from normality appear in larger sample sizes as a result of convergence to a normal distribution not being uniform in $\psi$. By the same token, the improved error-adjusted estimator is more biased than the error-adjusted estimator in larger samples, and even than the standard IV estimator in some scenarios. Informally, this happens because data sets which carry evidence for causal effects close to zero, yield estimated probabilities of $\hat{\delta}(\psi) \in \Delta$ close to zero. The bias then arises because the small estimated causal effects in such data sets will be more attracted towards the estimates obtained from a standard structural mean analysis (which ignores measurement error) than large estimated causal effects. Additional simulations (not displayed) have shown that, as predicted by the theory, this bias and deviation from normality disappears again in larger sample sizes. Furthermore, note that the confidence intervals for the improved error-adjusted estimator retain their coverage despite these deviations, although there is a tendency for the approach to be conservative. Finally, as predicted by the theory, the uniform confidence intervals are conservative and also wider on average than those obtained via the improved error-adjusted estimator. The impact of narrower intervals $\Delta = [-0.25, 0.25]$ was large at small sample sizes, but moderate at large sample sizes. For instance, confidence intervals based on the improved error-adjusted estimator had an average length of 8.42 (instead of 13.3) and coverage of 97.0% (instead of 97.7%) in the small samples and 4.35 (instead of 4.83) and 98.0% (instead of 97.8%), respectively, in the large samples. The impact of $\Delta = [-0.05, 0.05]$ not including the error mean was to induce bias of the order of magnitude of the standard unadjusted estimator. The 95% confidence intervals based on the improved error-adjusted estimator and uniform 95% confidence intervals then no longer cover at the nominal rate. Coverage of those intervals was still better than the coverage

Table 4.1: *Bias of the different effect estimators and coverage and average length of corresponding 95% confidence intervals.*

| $\Delta$ | $n$ | $\psi$ | $\delta$ | Bias | | | | Coverage | | | | | Average length CI | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | STD | IV1 | IV2 | IV3 | STD | IV1 | IV2 | IV3 | UI | STD | IV1 | IV2 | IV3 | UI |
| 0.5 | 105 | -7.5 | 0.15 | 1.11 | -3.77 | -2.65 | 0.68 | 86.8 | 96.5 | 99.8 | 97.7 | 99.8 | 5.87 | 3039 | 56.2 | 13.3 | 18.8 |
| 0.5 | 105 | -7.5 | 0 | -0.020 | -3.77 | -2.31 | -0.046 | 93.7 | 96.5 | 99.9 | 98.7 | 100 | 6.96 | 3039 | 55.7 | 10.8 | 29.2 |
| 0.5 | 105 | 0 | 0 | -0.015 | -3.63 | -0.019 | -0.0096 | 93.5 | 96.5 | 100 | 94.1 | 96.2 | 6.94 | 3027 | 63.8 | 8.88 | 21.1 |
| 0.5 | 1000 | -7.5 | 0.15 | 1.13 | -0.15 | -0.28 | 0.81 | 36.4 | 95.1 | 99.9 | 97.8 | 100 | 1.90 | 14.0 | 14.0 | 4.83 | 11.4 |
| 0.5 | 1000 | -7.5 | 0 | -0.0048 | -0.15 | -0.52 | -0.62 | 95.0 | 95.1 | 98.1 | 95.1 | 100 | 2.25 | 14.0 | 14.0 | 12.4 | 16.7 |
| 0.5 | 1000 | 0 | 0 | -0.0042 | -0.14 | 0.0032 | -0.0036 | 94.9 | 95.2 | 100 | 95.0 | 96.2 | 2.24 | 14.0 | 13.9 | 2.78 | 5.86 |
| 0.25 | 105 | -7.5 | 0.15 | 1.11 | -3.77 | 0.59 | 1.06 | 86.8 | 96.5 | 100 | 97.0 | 99.5 | 5.87 | 3039 | 56.4 | 8.42 | 10.9 |
| 0.25 | 105 | -7.5 | 0 | -0.020 | -3.77 | -0.63 | -0.062 | 93.7 | 96.5 | 100 | 98.8 | 99.9 | 6.96 | 3039 | 56.8 | 10.7 | 14.2 |
| 0.25 | 105 | 0 | 0 | -0.015 | -3.63 | -0.013 | -0.010 | 93.5 | 96.5 | 100 | 94.2 | 94.9 | 6.94 | 3027 | 67.8 | 8.88 | 11.6 |
| 0.25 | 1000 | -7.5 | 0.15 | 1.13 | -0.15 | 0.42 | 0.78 | 36.4 | 95.1 | 100 | 98.0 | 99.9 | 1.90 | 14.0 | 13.9 | 4.35 | 5.69 |
| 0.25 | 1000 | -7.5 | 0 | -0.0048 | -0.15 | -0.37 | -0.19 | 95.0 | 95.1 | 100 | 95.8 | 99.6 | 2.25 | 14.0 | 14.0 | 5.83 | 7.61 |
| 0.25 | 1000 | 0 | 0 | -0.0042 | -0.14 | -0.0012 | -0.0036 | 94.9 | 95.2 | 100 | 95.0 | 95.5 | 2.24 | 14.0 | 13.9 | 2.78 | 3.43 |
| 0.05 | 105 | -7.5 | 0.15 | 1.11 | -3.77 | 1.07 | 1.11 | 86.8 | 96.5 | 100 | 91.3 | 94.8 | 5.87 | 3039 | 63.9 | 6.54 | 7.47 |
| 0.05 | 105 | -7.5 | 0 | -0.020 | -3.77 | -0.052 | -0.015 | 93.7 | 96.5 | 100 | 96.3 | 98.2 | 6.96 | 3039 | 64.4 | 7.87 | 9.07 |
| 0.05 | 105 | 0 | 0 | -0.015 | -3.63 | -0.014 | -0.016 | 93.5 | 96.5 | 100 | 94.1 | 94.3 | 6.94 | 3027 | 67.9 | 7.53 | 8.51 |
| 0.05 | 1000 | -7.5 | 0.15 | 1.13 | -0.15 | 1.02 | 1.11 | 36.4 | 95.1 | 100 | 54.2 | 74.9 | 1.90 | 14.0 | 13.9 | 2.32 | 2.78 |
| 0.05 | 1000 | -7.5 | 0 | -0.0048 | -0.15 | -0.031 | -0.0063 | 95.0 | 95.1 | 100 | 98.3 | 99.7 | 2.25 | 14.0 | 13.9 | 2.83 | 3.43 |
| 0.05 | 1000 | 0 | 0 | -0.0042 | -0.14 | -0.0037 | -0.0042 | 94.9 | 95.2 | 100 | 95.0 | 95.1 | 2.24 | 14.0 | 13.9 | 2.36 | 2.55 |

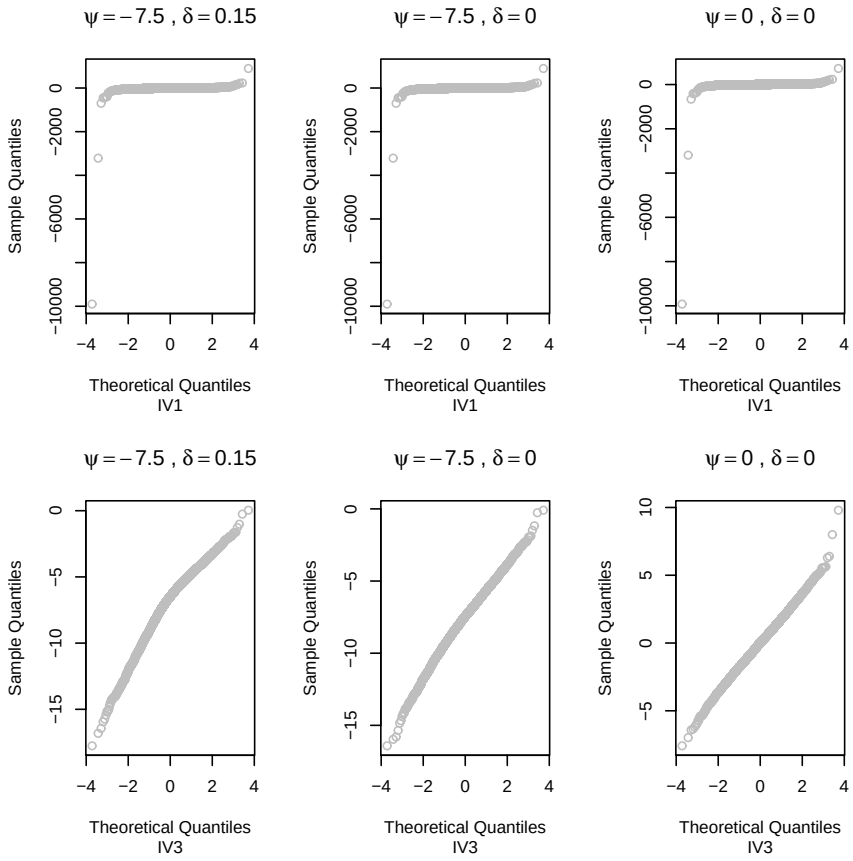



Figure 3: *QQ-plots for* $n = 105$ *and* $\Delta = [-0.5, 0.5]$. *Row 1: error-adjusted estimator IV1; Row 2: improved error-adjusted estimator IV3.*

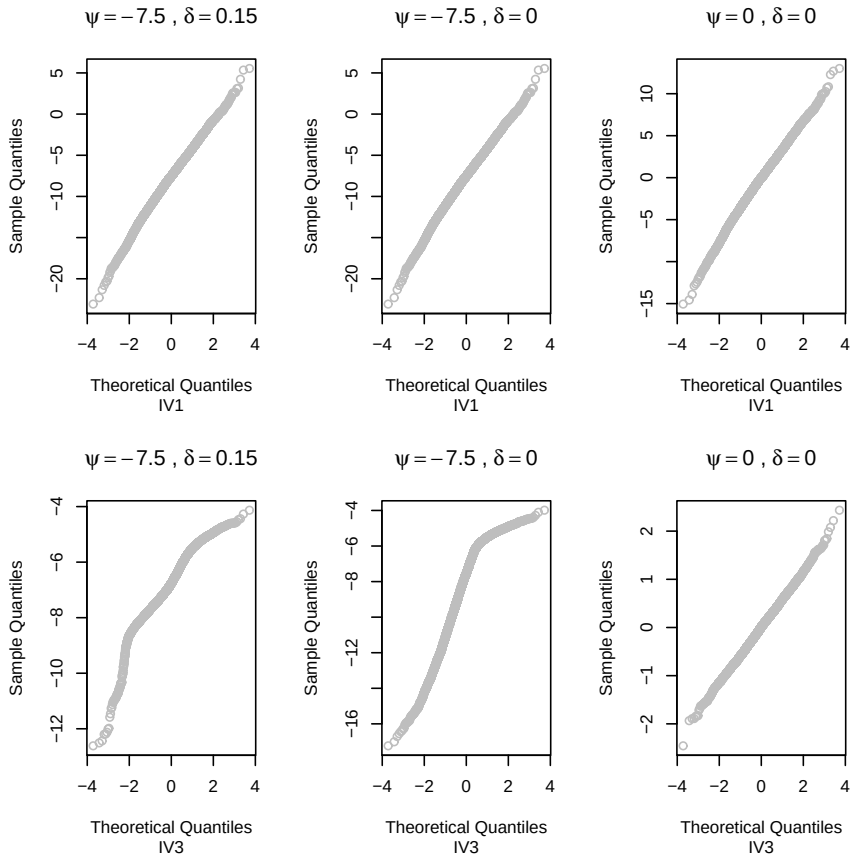



Figure 4: *QQ-plots for* $n = 1000$ *and* $\Delta = [-0.5, 0.5]$. *Row 1: error-adjusted estimator IV1; Row 2: improved error-adjusted estimator IV3.*

of 95% confidence intervals based on the standard unadjusted estimator, but at the expense of being wider.

# 6 Conclusions

We have proposed a general procedure to correct IV estimators for systematic error in the exposure when an additional instrumental variable for the measurement error is available. This procedure complements the sensitivity analysis approach of Goetghebeur and Vansteelandt (2005) and is especially attractive when the IV-M assumption (A4) is likely to be met. This is the case in placebo-controlled randomized trials with noncompliance where measurements $\mathbf{T}_i$ on run-in placebo compliance may very well meet assumption (A4). With concern for compliance mismeasurement, recording run-in compliance may thus be favourable. More generally, IV-C's can be used as IV's for the measurement error.

On theoretical grounds and on the basis of simulation experiments, we recommend the improved error-adjusted estimator of Section 3.1. This estimator was designed so that adjustment for measurement error does not compromise the power of tests of the causal null. This is attractive, knowing that standard tests of the causal null hypothesis (i.e., that the causal instrument $R$ is independent of outcome) ignore exposure measurements and are thus valid in the presence of measurement error. Because the proposed estimator does not converge uniformly to a normal distribution, we recommend the uniform confidence intervals of Section 3.2.

For illustrative purposes, we have developed this work under structural mean models which assume linear exposure effects that are not modified by pre-exposure covariates. Extensions to linear structural mean models that allow for effect modification by baseline covariates are methodologically straightforward, but computationally more demanding. Finally, we believe our results to be more broadly useful as they suggest, in line with Gustafson (2005), that incorporating a little prior information on a weakly identified nuisance parameter may yield substantial efficiency improvements for the target parameter. Similar ideas may therefore prove useful in related settings (Vansteelandt and Goetghebeur, 2004; Fischer and Goetghebeur, 2004; Ten Have et al., 2007) with weak identification. In addition, our results indicate how such prior information may be adopted in a frequentist analysis.

# Appendix 4.A: Proof of Proposition 1 and Theorems 1, 2

**Proof of Proposition 1.** Model $\mathcal{A}$ implies model $\mathcal{B}$ because, with

$$\delta(\mathbf{X}_i, R_i) \equiv E(W_i - Z_i|\mathbf{X}_i, R_i)$$

,

$$\begin{aligned} E\left(Y_{i0}|\mathbf{X}_i, R_i\right) &= E\left\{Y_i - \gamma(\mathbf{S}_i; \psi^*)Z_i|\mathbf{X}_i, R_i\right\} \\ &= E\left[Y_i - \gamma(\mathbf{S}_i; \psi^*)\left\{W_i - \delta(\mathbf{X}_i, R_i)\right\}|\mathbf{X}_i, R_i\right] \end{aligned}$$

by (A3) and (A5), and because $E\left(Y_{i0}|\mathbf{X}_i, R_i\right) = E\left(Y_{i0}|\mathbf{X}_i\right)$ by (A1). Note that this does not require assumptions about the conditional association between $Y_i$ and $W_i$, given $Z_i$, suggesting that this continues to hold when measurement error is differential.

To show that (4.6) is the only restriction (other than (4.5)) imposed on the observed data law, we proceed as in Robins and Rotnitzky (2004) by exhibiting for any observed data law satisfying (4.5) and (4.6), a joint law of the full data $(Y, \{Y_{rz}, \forall r, z\}, Z, W, R, \mathbf{X})$ satisfying the restrictions of model $\mathcal{A}$, where $Y_{rz}$ is the potential outcome that would have been observed for given subject following exposure to $(R, Z) = (r, z)$, all other experimental conditions being the same as in the considered study. Given $(R = r, Z = z, W = w, \mathbf{X} = \mathbf{x}, Y = y)$, we define $Y_{rz} = y$ to satisfy (A2). We set $f(Z|R = r, W = w, \mathbf{X} = \mathbf{x}, Y = y)$ equal to an arbitrary density with conditional mean $w - \delta$ to satisfy (A5). We define $f(Y_{r0}|R = r, Z = z, W = w, \mathbf{X} = \mathbf{x}, Y = y)$ to be an arbitrary density with conditional mean $y - \gamma(\mathbf{x}, r; \psi^*)z$. In addition, given $(R = r, Z = z, W = w, \mathbf{X} = \mathbf{x}, Y = y)$,

we set $Y_{r0} = Y_{r'0} \equiv Y_0$ for each $r'$ to satisfy (A1). By (4.6), the conditional distribution of $Y_0$ then also satisfies $E(Y_0|\mathbf{X} = \mathbf{x}, R) = E(Y_0|\mathbf{X} = \mathbf{x})$ for each $\mathbf{x}$. Remaining features of the full data density can be chosen arbitrarily.

**Proof of Theorem 1.** Let for simplicity of exposition, but without loss of generality, $\gamma(\mathbf{X}_i; \psi) = \psi, Z_i = Z_i R_i$ and $E(W_i - Z_i|\mathbf{X}_i, R_i) = \delta^* R_i$. Define $U_{i\delta} = d_\delta(R_i, \mathbf{X}_i)\left[Y_i - \psi(W_i - \delta)R_i - q(\mathbf{X}_i)\right]$ and $U_{i\psi} = d_\psi(R_i, \mathbf{X}_i)\left[Y_i - \psi(W_i - \delta)R_i - q(\mathbf{X}_i)\right]$ the estimating functions for $\delta^*$ and $\psi^*$, respectively. Under weak regularity conditions as stated for general M-estimators in van der Vaart (1998, p.48, 60), Taylor expansions show that

$$\begin{aligned} 0 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_{i\delta} + E\left(\frac{\partial U_\delta}{\partial \psi}\right)\sqrt{n}(\hat{\psi} - \psi^*) + E\left(\frac{\partial U_\delta}{\partial \delta}\right)\sqrt{n}(\hat{\delta} - \delta^*) \\ &\quad + \frac{1}{2}E\left(\frac{\partial^2 U_\delta}{\partial \psi \partial \delta}\right)\sqrt{n}(\hat{\psi} - \psi^*)(\hat{\delta} - \delta^*) + o_p(1) \end{aligned} \tag{4.11}$$

from which

$$\begin{aligned} \sqrt{n}(\hat{\delta} - \delta^*)\frac{\hat{\psi} + \psi^*}{2} &= o_p(1) - E^{-1}\left\{d_\delta(R, \mathbf{X})R\right\} \\ &\quad \times \left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_{i\delta} - E\left\{d_\delta(R, \mathbf{X})(W - \delta^*)R\right\}\sqrt{n}(\hat{\psi} - \psi^*)\right] \end{aligned}$$

Plugging this into a first order Taylor expansion of $U_{i\psi}$, shows that

$$\begin{aligned} \sqrt{n}(\hat{\psi} - \psi^*) = &- \left[E\left\{d_\psi(R, \mathbf{X})(W - \delta^*)R\right\} - \frac{E\left\{d_\psi(R, \mathbf{X})R\right\}}{E\left\{d_\delta(R, \mathbf{X})R\right\}}E\left\{d_\delta(R, \mathbf{X})(W - \delta)R\right\}\right]^{-1} \\ &\times \left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_{i\psi} - \frac{E\left\{d_\psi(R, \mathbf{X})R\right\}}{E\left\{d_\delta(R, \mathbf{X})R\right\}}U_{i\delta}\right] + o_p(1) \end{aligned}$$

Standard application of the Central Limit Theorem and Slutsky's Theorem now shows that $\sqrt{n}(\hat{\psi} - \psi^*) = O_p(1)$ and that Part 1 of Theorem 1 holds.

Note that the last 3 terms in (4.11) can be replaced with

$$E\left\{d_\delta(R, \mathbf{X})R\right\}\left\{\psi^* + O_p(n^{-1/2})\right\}\sqrt{n}(\hat{\delta} - \delta),$$

from which $\sqrt{n}(\hat{\delta}-\delta)\psi^* = \sqrt{n}(\hat{\delta}-\delta)(\hat{\psi}+\psi^*)\{1/2+o_p(1)\}$ equals

$$-\left[E\{d_\delta(R,\mathbf{X})R\} - \frac{E\{d_\delta(R,\mathbf{X})(W-\delta^*)R\}}{E\{d_\psi(R,\mathbf{X})(W-\delta^*)R\}}E\{d_\psi(R,\mathbf{X})R\}\right]^{-1}$$
$$\times\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}U_{i\delta} - \frac{E\{d_\delta(R,\mathbf{X})(W-\delta^*)R\}}{E\{d_\psi(R,\mathbf{X})(W-\delta^*)R\}}U_{i\psi}\right] + o_p(1)$$

The latter expression is bounded in probability (under standard regularity conditions). It follows that, as $\psi^*$ goes to zero with increasing sample size, $\hat{\delta}$ does not converge to $\delta^*$ at root-$n$ rate and hence is not uniformly root-$n$ consistent. In particular, there is no root-$n$ consistent estimator of $\delta^*$ under model $\mathcal{A}$ at $\psi^*=0$, which proves Part 2 of Theorem 1. This is also seen by noting that the expected derivative of the efficient estimating function for $(\psi,\delta)$ w.r.t. $\delta$ is zero at $\psi=0$.

Part 3 of Theorem 1 is immediate from Robins (1994).

**Proof of Theorem 2.** Let for simplicity of exposition, but without loss of generality, $\gamma(\mathbf{X}_i;\psi)=\psi, Z_i = Z_iR_i$ and $E(W_i - Z_i|\mathbf{X}_i,R_i)=\delta^*R_i$. Then standard asymptotic theory for M-estimators (van der Vaart, 1998) and Taylor expansions of the estimating functions (4.9) for $\psi^*$ w.r.t. $\hat{\delta}(\psi)$ shows that (4.9) equals

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}P\{\hat{\delta}(\psi)\in\Delta\}U_{i\psi}(\psi,\delta) + P\{\hat{\delta}(\psi)\notin\Delta\}U_{i\psi}(\psi,0) + o_p(1) - \Big[P\{\hat{\delta}(\psi)\in\Delta\}$$
$$+\left\{\varphi\left(\frac{\Delta_l-\delta}{\sigma(\psi)/(\sqrt{n}|\psi|)}\right) - \varphi\left(\frac{\Delta_u-\delta}{\sigma(\psi)/(\sqrt{n}|\psi|)}\right)\right\}\frac{\sqrt{n}|\psi|\delta}{\sigma(\psi)}\Big]\frac{E\{d_\psi(R,T)R\}}{E\{d_\delta(R,T)R\}}U_{i\delta}(\psi,\delta) \quad (4.12)$$

That the remainder term converges to zero in probability for any fixed $\psi$ can be seen because, for some $\tilde{\delta}$ on the open line segment between $\hat{\delta}(\psi)$ and $\delta^*$ (under standard regularity conditions which include uniform convergence of $n^{-1}\sum_{i=1}^{n}U_{i\psi}(\psi,\delta)$ w.r.t. $\delta$), the remainder term equals

$$\left[P_{\delta=\tilde{\delta}}\{\hat{\delta}(\psi)\in\Delta\}E\left\{\frac{\partial^2}{\partial\delta^2}U_{i\psi}(\psi,\tilde{\delta})\right\} + 2\frac{\partial}{\partial\delta}P_{\delta=\tilde{\delta}}\{\hat{\delta}(\psi)\in\Delta\}E\left\{\frac{\partial}{\partial\delta}U_{i\psi}(\psi,\tilde{\delta})\right\}\right.$$
$$\left.+\frac{\partial^2}{\partial\delta^2}P_{\delta=\tilde{\delta}}\{\hat{\delta}(\psi)\in\Delta\}E\left\{U_{i0}(\psi) - U_{i\psi}(\psi,\tilde{\delta})\right\}\right]\frac{\sqrt{n}}{2}\{\hat{\delta}(\psi)-\delta^*\}^2 + o_p(1)$$

Here, $E\left\{\partial^2U_{i\psi}(\psi,\tilde{\delta})/\partial\delta^2\right\}=0$. Because $E\left\{\partial U_{i\psi}(\psi,\tilde{\delta})/\partial\delta\right\} = O_p(1)\psi$ under standard regularity conditions and $\sqrt{n}\{\hat{\delta}(\psi)-\delta^*\}^2 = O_p(1)n^{-1/2}\psi^{-2}$, the second term

is

$$O_p(1)\left\{\varphi\left(\frac{\Delta_l-\tilde{\delta}}{\sigma(\psi)/(\sqrt{n}|\psi|)}\right)-\varphi\left(\frac{\Delta_u-\tilde{\delta}}{\sigma(\psi)/(\sqrt{n}|\psi|)}\right)\right\}\frac{1}{\sigma(\psi)}=o_p(1)$$

for any fixed $\psi$. Because $E\left\{U_{i0}(\psi)-U_{i\psi}(\psi,\tilde{\delta})\right\}=O_p(1)\tilde{\delta}\psi$, the third term is

$$\left\{\varphi\left(\frac{\Delta_l-\tilde{\delta}}{\sigma(\psi)/(\sqrt{n}|\psi|)}\right)-\varphi\left(\frac{\Delta_u-\tilde{\delta}}{\sigma(\psi)/(\sqrt{n}|\psi|)}\right)\right\}\frac{n|\psi|\psi\tilde{\delta}}{\sigma(\psi)^3}(\Delta_l-\Delta_u)=o_p(1)$$

for any fixed $\psi$ because $x^a\varphi(x)\to 0$ as $x\to\infty$ for arbitrary $a>0$.

Because the estimating functions in (4.12) have mean and variance depending on the sample size, we use the triangular array Central Limit Theorem (Serfling, 1980, p.31) to derive the asymptotic distribution of (4.9) for fixed $\psi$. Application of this Theorem shows that for arbitrary fixed $\psi$, the estimating functions in (4.9) are asymptotically normally distributed under the weak regularity condition that the standard deviation of the estimating functions $\tilde{U}_i(\psi)$, as defined by (4.12), is bounded (i.e. $O(1)$) and that asymptotically $E\|\tilde{U}_i(\psi)-E\{\tilde{U}_i(\psi)\}\|^k=o(n^{k/2-1})$ for each $k$. Because for any fixed $\psi^*\neq 0$ and $\delta^*\in\Delta=]\Delta_l,\Delta_u[$, $P\{\hat{\delta}(\psi^*)\in\Delta\}$ converges to 1, it follows under these conditions that $n^{-1/2}\sum_{i=1}^n\tilde{U}_i(\psi^*)$ will be asymptotically normally distributed with mean zero and finite variance, which is given by the variance of (4.12). Within faster than root-$n$ shrinking neighbourhoods of zero (i.e. if $\psi^*=kn^{-a}$ for some constant $k$ and $a>1/2$), the remainder term in the Taylor series expansion is still $o_p(1)$. Further, $P\{\hat{\delta}(\psi^*)\in\Delta\}$ converges to 0 and $U_0(\psi^*)$ has mean converging to zero at 1 over $n^a$-rate. It then again follows that $n^{-1/2}\sum_{i=1}^n\tilde{U}_i(\psi^*)$ is asymptotically normally distributed with mean zero and finite variance. Finally, within 1 over root-$n$ shrinking neighbourhoods of zero (i.e. if $\psi^*=kn^{-1/2}$ for some constant $k$), the remainder term in the Taylor series expansion is bounded in probability, but not $o_p(1)$. The significant contribution of the squared term $\sqrt{n}\{\hat{\delta}(\psi^*)-\delta^*\}^2$ implies that $n^{-1/2}\sum_{i=1}^n\tilde{U}_i(\psi^*)$ may not converge to a normal distribution, nor to a mean zero distribution along such sequences. The implications of this will be discussed in the next paragraph.

To gain insight into the asymptotic distribution of $\tilde{\psi}$ (rather than its estimating function), we make a further Taylor series expansion of the estimating functions,

evaluated at $\tilde{\psi}$. This shows that for any fixed $\psi$

$$
\begin{aligned}
0 = {} & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} P\{\hat{\delta}(\psi) \in \Delta\} U_{i\psi}(\psi, \delta) + P\{\hat{\delta}(\psi) \notin \Delta\} U_{i\psi}(\psi, 0) + o_p(1) - \Bigg[ P\{\hat{\delta}(\psi) \in \Delta\} \\
& + \left\{ \varphi\left( \frac{\Delta_l - \delta}{\sigma(\psi)/(\sqrt{n}|\psi|)} \right) - \varphi\left( \frac{\Delta_u - \delta}{\sigma(\psi)/(\sqrt{n}|\psi|)} \right) \right\} \frac{\sqrt{n}|\psi|\delta}{\sigma(\psi)} \Bigg] \frac{E\{d_\psi(R,X)R\}}{E\{d_\delta(R,X)R\}} U_{i\delta}(\psi, \delta) \\
& + \left( P\{\hat{\delta}(\psi) \in \Delta\} E\left\{ \frac{\partial}{\partial \psi} U_\psi(\psi, \delta) \right\} + P\{\hat{\delta}(\psi) \notin \Delta\} E\left\{ \frac{\partial}{\partial \psi} U_\psi(\psi, 0) \right\} \right. \\
& + \left\{ \varphi\left( \frac{\Delta_l - \delta}{\sigma(\psi)/(\sqrt{n}|\psi|)} \right) - \varphi\left( \frac{\Delta_u - \delta}{\sigma(\psi)/(\sqrt{n}|\psi|)} \right) \right\} \frac{\sqrt{n}|\psi|\delta(\Delta_l - \Delta_u)}{\sigma(\psi)} E\{d_\psi(R,X)R\} \\
& - \left[ P\{\hat{\delta}(\psi) \in \Delta\} + \left\{ \varphi\left( \frac{\Delta_l - \delta}{\sigma(\psi)/(\sqrt{n}|\psi|)} \right) - \varphi\left( \frac{\Delta_u - \delta}{\sigma(\psi)/(\sqrt{n}|\psi|)} \right) \right\} \frac{\sqrt{n}|\psi|\delta}{\sigma(\psi)} \right] \\
& \left. \times \frac{E\{d_\psi(R,X)R\}}{E\{d_\delta(R,X)R\}} E\{d_\delta(R,X)R(W-\delta)\} \right) \sqrt{n}(\tilde{\psi} - \psi)
\end{aligned} \tag{4.13}
$$

That the remainder term converges to zero in probability for any fixed $\psi$ can be seen using a similar derivation as before. We conclude that, up to an $o_p(1)$ term and for fixed $\psi$, $\sqrt{n}(\tilde{\psi} - \psi)$ is a linear transformation of $n^{-1/2} \sum_{i=1}^{n} \tilde{U}_i(\psi)$ and thus shares its asymptotic properties. Specifically, within faster and slower than 1 over root-$n$ shrinking neighbourhoods of zero (and in particular at arbitrary fixed $\psi$), $\sqrt{n}(\tilde{\psi} - \psi)$ is asymptotically normally distributed with mean zero and finite variance under weak regularity conditions. Within 1 over root-$n$ neighbourhoods of zero, $\sqrt{n}(\tilde{\psi} - \psi)$ may be asymptotically biased and not normally distributed.

# Chapter 5

# Comparison of Causal Effect Estimators Under Exposure Misclassification


## Summary

Over the past decades, various principles for causal effect estimation have been proposed, all differing in terms of how they adjust for measured confounders: either via traditional regression adjustment, by adjusting for the expected exposure given those confounders (e.g., the propensity score), or by inversely weighting each subject's data by the probability of the observed exposure, given those confounders. When the exposure is measured with error, this raises the question whether these different estimation strategies might be differently affected under the assumption of no unmeasured confounders. In this article, we investigate this by comparing inverse probability of treatment weighted (IPTW) estimators and doubly robust estimators for the exposure effect in linear marginal structural mean models (MSM) with G-estimators, propensity score (PS) adjusted estimators and ordinary least squares (OLS) estimators for the exposure effect in linear regression models. We find analytically that these estimators are equally affected when exposure misclassification is independent of the confounders, but not otherwise. Simulation studies reveal similar results for time-varying exposures and when the model of interest includes a logistic link.

# 1 Introduction

Measurement error in exposures is inevitable in many observational studies, in particular when exposures are difficult to measure (e.g. individual exposure to pollutants) or when risk behaviours such as condom use, diet, smoking, alcohol consumption, etc. are obtained through self-report or questionnaires (Buonaccorsi et al., 2005; Landin et al., 1995; Spiegelman et al., 2001). It forms a common source of bias in estimates of exposure effects, in particular when those effects are confounded by measured covariates (Budtz-Jørgensen et al., 2003). In view of this, a large literature has come available on how to quantify the effects of exposure mismeasurement on adjusted regression coefficients (Buonaccorsi et al., 2005; Budtz-Jørgensen et al., 2003; Veierod and Laake, 2001), and on how to correct for this when an accurate estimate of the error variance or misclassification probability is known (Gustafson et al., 2001; Carroll et al., 2006) or when sources of information on the measurement error (e.g. from validation studies) are available (Lyles et al., 2007; Thürigen et al., 2000).
Over the past decades, alternatives to ordinary regression adjustment for confounder control have been proposed, based on adjustment for the propensity score (i.e. the probability of being exposed, given measured background characteristics) (Rosenbaum and Rubin, 1983; Rubin, 1997), and on inverse weighting by the probability of the observed exposure, given measured background characteristics (Robins et al., 2000). These strategies have mainly been introduced:

- because they work better in settings where there is little overlap in the distribution of confounders between exposed and unexposed subjects (Rubin, 1997; Lunceford and Davidian, 2004);
- because some allow correction for confounders which are intermediate on the path from early exposure to outcome (Robins et al., 2000);
- and because they separate the modelling of confounders from the modelling of treatment effects, thus being less vulnerable to data-driven presentation of the most favorable treatment effect estimators (Moore and van der Laan, 2007).

To the best of our knowledge, the impact of exposure mismeasurement on such estimators has not previously been investigated. The purpose of this paper is therefore to explore and illustrate the consequences of exposure mismeasurement on these different methods for confounder control. Our interest in this is partly motivated by the fact that some of the considered methods not only use the exposure as a covariate in a regression model, but also rely on estimates of the expected exposure, given background covariates, and may therefore be differently affected by measurement error. Throughout the paper, our focus will be on misclassification of dichotomous

exposures because propensity score and IPTW methods were specifically designed to analyze such exposures. We do not consider instrumental variables estimators on which we report elsewhere (Vansteelandt and Goetghebeur, 2005; Vansteelandt et al., 2007).

## 2 Estimating the causal effect of point-exposures

Consider a study whose goal is to estimate the causal effect of a binary exposure (or dose) $D$ ($D = 1$ if exposed, 0 if not) based on a random population sample. Let $Y_d$ represent the potential outcome which a subject would have had if, possibly contrary to fact, the exposure $D$ were set to $d$ (Rubin, 1974). The expected contrast $\beta^* = \mathrm{E}(Y_1 - Y_0)$ then defines the average causal effect of exposure on outcome. Assume that for each subject we either obtain $Y_0 = Y$ (when $D = 0$) or $Y_1 = Y$ (when $D = 1$). Suppose furthermore that all confounders $X$ for the association between $D$ and $Y$ have been accurately measured, so that the no unmeasured confounders assumption

$$D \perp\!\!\!\perp Y_d | X \tag{5.1}$$

for $d = 0, 1$ holds. Then $\beta^*$ may be evaluated under the following semiparametric model,

$$\mathrm{E}(Y_d | X) = \beta^* d + g(X) \tag{5.2}$$

where $g(X)$ is a unknown function of the covariates $X$. Note that this model has a parametric component $\beta^* d$ and a nonparametric component $g(X)$ (Robins et al., 1992). Alternatively, $\beta^*$ can be estimated by fitting the marginal structural mean model (MSM),

$$\mathrm{E}(Y_d) = \beta^* d + \alpha^* \tag{5.3}$$

where $\alpha^* = \mathrm{E}\{g(X)\}$. Note that model (5.2) is more restrictive than model (5.3) because it assumes that the exposure effect is not modified by the given covariates $X$. Model (5.3) may therefore be of greater interest in settings where there is no specific interest in effect modification.

Inference for $\beta^*$ in models (5.2) and (5.3) has been extensively described (see e.g. Robins et al., 2000 and 1992) and different estimators for $\beta^*$ in these models have been proposed. Due to curse of dimensionality, these estimators rely on correct specification of certain working models:

- Some postulate a model for the conditional association $g(X)$ between $Y$ and $X$, given $D$. For example, standard OLS estimators are obtained by assuming a parametric working model $g(X) = \alpha^* + \delta^* X$ (with $(\alpha^*, \delta^*)$ unknown) at the risk

of finding biased causal effect estimators when this model is misspecified and at the risk of extrapolation when the distribution of $X$ has little overlap between exposed and unexposed subjects (Rubin, 1997; Lunceford and Davidian, 2004).

- Some rely on a model for the marginal distribution of $D$, given $X$ (e.g. the IPTW estimator of Section 2.1). As such, they may avoid explicit regression extrapolation and be more reliable in settings where the investigator has better a priori knowledge of the association between exposure and confounders, than about the association between outcome and confounders.

- Finally, so-called doubly robust estimators of $\beta^*$ assume correct specification of at least one of these 2 working models, but not necessarily both (e.g. the doubly robust IPTW estimator of Section 2.1 or the G-estimator of Section 2.2). The latter estimators are therefore more robust to misspecification of the working models.

In the following sections, we briefly review these different estimators, assuming that we have available a random sample of measurements $(Y_i, D_i, X_i)$ for independent subjects $i = 1, ..., n$. In Section 3, we then examine the impact of exposure misclassification on these estimators.

## 2.1 (Doubly robust) IPTW estimation

Estimates for the parameter $\beta^*$ in model (5.3) can be obtained by IPTW estimation; that is, by fitting a linear regression model for $Y$ on $D$ where the contribution of each subject $i$ is weighted by 1 over the probability of that subject's observed exposure given the confounders $X$. Indeed, the impact of inverse weighting by the probability $P(D|X)$ is to remove the association between $D$ and $X$ (while leaving the causal effect unchanged), thus eliminating the need to control for $X$. IPTW estimation under model (5.3) thus proceeds by solving the weighted regression equation

$$\sum_{i=1}^{n} U_{iI}(\alpha, \beta) = \sum_{i=1}^{n} \begin{pmatrix} 1 \\ D_i \end{pmatrix} \frac{1}{P(D_i|X_i)} (Y_i - \alpha - \beta D_i) = 0 \tag{5.4}$$

for $(\alpha, \beta)$, where index I stands for IPTW estimator. In practice, the probabilities $P(D_i|X_i)$ are unknown and must be estimated. Because $D$ is a dichotomous exposure, this may be done by fitting the logistic regression model

$$P(D_i = 1|X_i; \theta) = \text{expit}\,(\theta_0 + \theta_1 X_i) \tag{5.5}$$

for $P(D_i = 1|X_i)$, where $\text{expit}(z) = e^z/(1+e^z)$. Estimation thus proceeds in two stages: first estimate $\theta$ in model (5.5); next, solve (5.4) using the fitted values from model (5.5) to obtain an estimate for $\beta^*$. A drawback of IPTW estimators is that they can be quite unstable and inefficient when confounding is severe. This is because the probabilities $P(D_i|X_i)$ are then typically small for some individuals, who thus receive a large weight in the analysis, and may become influential. Doubly robust (DR) estimators (van der Laan and Robins, 2002; Yu and van der Laan, 2003) alleviate this problem by allowing misspecification of the inverse weights $1/P(D_i|X_i)$, provided that a conditional mean model for $\text{E}(Y_i|D_i, X_i)$ holds. As a result, they allow truncation of extreme weights (Yu and van der Laan, 2003). These estimators have the added advantage of being efficient under the model defined by (5.3) and (5.5) when in addition to these models, the conditional mean $\text{E}(Y_i|D_i, X_i)$ is correctly specified.

In practice, doubly robust estimator can be obtained by solving an estimating equation of the form

$$\begin{aligned}\sum_{i=0}^{n} U_{iD}(\alpha,\beta) &= \sum_{i=1}^{n}\Bigg[\begin{pmatrix}1\\ D_i\end{pmatrix}\frac{Y_i - E(Y_i|D_i,X_i)}{P(D_i|X_i)} \\ &\quad + \sum_{d=0}^{1}\begin{pmatrix}1\\ d\end{pmatrix}\{E(Y_i|D_i=d,X_i) - \alpha - \beta d\}\Bigg] = 0 \end{aligned} \tag{5.6}$$

where index D stands for doubly robust estimator. Here $P(D_i|X_i)$ and $\text{E}(Y_i|D_i, X_i)$ are unknown, but can be estimated by fitting a model $P(D_i|X_i;\theta)$ for $P(D_i|X_i)$ as in the previous paragraph and a model $\text{E}(Y_i|D_i, X_i;\gamma)$ for $\text{E}(Y_i|D_i, X_i)$, for example

$$\text{E}(Y_i|D_i, X_i;\gamma) = \gamma_0 + \gamma_1 D_i + \gamma_2 X_i. \tag{5.7}$$

Estimation then proceeds in three stages: first obtain an estimate $\hat{\theta}$ of $\theta$ in model (5.5); then, obtain an estimate $\hat{\gamma}$ of $\gamma$ in model (5.7); finally, solve (5.6) using fitted values from models (5.5) and (5.7) to obtain an estimate

$$\hat{\beta}_D = \frac{1}{n}\sum_{i=1}^{n}(2D_i - 1)\frac{Y_i - \text{E}(Y_i|D_i, X_i;\hat{\gamma})}{P(D_i|X_i;\hat{\theta})} + \hat{\gamma_1} \tag{5.8}$$

for $\beta^*$. The above methods continue to work for logistic MSMs for a binary outcome

$$P(Y_d = 1) = \text{expit}\,(\alpha^* + \beta^* d) \tag{5.9}$$

upon substituting $\alpha^* + \beta^* d$ in the above estimating equations with $\text{expit}(\alpha^* + \beta^* d)$. Note that, due to noncollapsibility of the odds ratio, the parameter $\beta^*$ in model (5.9)

differs from the corresponding parameter $\beta^{**}$ in model

$$P(Y_d = 1|X) = \text{expit}\left\{g(X) + \beta^{**}d\right\} \tag{5.10}$$

which can be estimated via traditional logistic regression under a parametric working model for $g(X)$. Because different studies typically adjust for different sets of covariates $X$, it follows that estimates of $\beta^*$ are better comparable between studies (and may thus be of greater interest) than estimates of $\beta^{**}$ (Moore and van der Laan, 2007).

## 2.2 G-estimation and propensity score adjustment

Estimates for the parameters $\beta^*$ indexing model (5.2) may be obtained via G-estimation (Robins et al., 1992; Brumback et al., 2003). The principle is that after subtracting the causal effect $\beta^* D$ from the outcome, the resulting 'treatment-free outcome' $Y - \beta^* D$ should be conditionally mean independent of $D$, given $X$, by the no unmeasured confounders assumption (5.1). One may thus estimate $\beta^*$ as the value for which this conditional independence holds in the observed data set; that is, by solving

$$\sum_{i=1}^{n} U_{iG}(\beta) = \sum_{i=1}^{n} \{D_i - E(D_i|X_i)\} \left[Y_i - \mathrm{E}(Y_i|X_i) - \beta \{D_i - E(D_i|X_i)\}\right] = 0 \tag{5.11}$$

where index G stands for G-estimator. This is equivalent to fitting the linear model

$$\mathrm{E}(Y_i|D_i, X_i; \gamma, \beta) = \gamma_0 + \gamma_1 X_i + \beta \{D_i - E(D_i|X_i)\} \tag{5.12}$$

or

$$\mathrm{E}(Y_i|D_i, X_i; \gamma, \beta) = \gamma_0 + \gamma_1 X_i + \gamma_2 E(D_i|X_i) + \beta D_i \tag{5.13}$$

and is thus equivalent to regression adjustment for the propensity score in linear models. This is because (5.11) is the projection of the score for $\beta$ in models (5.12) and (5.13) onto the orthocomplement for the tangent space of the parameters $\gamma_0, \gamma_1$ (and $\gamma_2$) (Bickel et al., 1993). Here, $E(D_i|X_i)$ is unknown, but can be estimated by fitting a model $E(D_i|X_i; \theta)$ for $E(D_i|X_i)$. Estimation thus proceeds in two stages: first estimate $\theta$ in model (5.5); then, fit model (5.12). The resulting estimator has the attractive property of being doubly robust in the sense that, $\hat{\beta}_G$ is a consistent estimator of $\beta^*$ when either working model $\mathrm{E}(D_i|X_i; \theta)$ or the model for $\mathrm{E}(Y_i|X_i)$ (i.e. $\mathrm{E}(Y_i|X_i) = \gamma_0^* + \gamma_1^* X_i$ for unknown parameters $\gamma_0^*, \gamma_1^*$ in model (5.12)) is correctly specified (Robins et al., 1992). G-estimation has no immediate extensions to logistic regression models for a dichotomous outcome (Robins et al., 1992), but traditional regression adjustment for the propensity score $P(D_i = 1|X_i)$ remains valid in such models.

# 3 The impact of point-exposure misclassification

In this section, we study the impact of exposure misclassification on the previous causal effect estimators. Specifically, suppose that instead of $D$, we observe an error-prone version $W$. Then our goal is to quantify the asymptotic bias due to exposure misclassification on each of the above effect estimators when the exposure misclassification probabilities are known. Intuitively, one would expect that the different estimators will be differently affected by measurement error. Indeed, OLS estimators merely rely on a conditional mean model for the outcome, which involves the error-prone exposure. In contrast, IPTW estimators may be affected by misclassification in 2 manners: (a) through estimation of the inverse weights; and (b) by substituting $W$ for $D$ in model (5.3) or model (5.9). Likewise, G-estimators and propensity score adjusted estimators may be affected by misclassification through estimation of the conditional probabilities $P(D|X)$, but could potentially be less affected than the IPTW estimators by avoiding inverse weighting.

## 3.1 Asymptotic bias

Throughout, we measure the degree of misclassification in terms of the conditional probabilities $\pi_{d|w,X} = P(D = d|W = w, X)$. Here, the probability $\pi_{1|1,X} = P(D = 1|W = 1, X)$ expresses how likely it is for someone who is classified as exposed with covariates level $X$ to be truly exposed, and likewise, $\pi_{0|0,X} = P(D = 0|W = 0, X)$ expresses how likely it is for someone who is classified as unexposed with covariates level $X$ to be truly unexposed. In line with Buonaccorsi et al. (2005), we call $\pi_{0|0,X}$ and $\pi_{1|1,X}$ reclassification probabilities. Note that it is more common to measure misclassification in terms of the probabilities $P(W = w|D = d, X)$ (Buonaccorsi et al., 2005; Veierod and Laake, 2001; Gustafson et al., 2001). Our reason not to adopt this more common definition is that IPTW estimators and G-estimators rely on a working model for $P(W|X)$. Given that $W$ and $X$ are jointly observed on each subject it is more natural to assume the model for $P(W|X)$ to be correctly specified and, in addition, to assume knowledge about the probabilities $P(D|W, X)$. Throughout this article, we assume non-differential misclassification of the exposure in the sense that among subjects with the same level of the measured confounder $X$, misclassification happens independently of the (potential) outcome; i.e. $Y_d \perp\!\!\!\perp W|D, X$ for $d = 0, 1$.

In the Appendix, we derive the bias of the (DR) IPTW estimator, the G-estimator and the OLS estimator due to misclassification under the linear models (5.2) and (5.3). This is done by calculating the limiting values $\alpha$ and $\beta$ that solve the expected (i.e. limiting) estimating equation corresponding to these estimators. For instance, for the

IPTW estimator, we solve

$$\mathrm{E}\{U_{iI}(\alpha,\beta)\} = \sum_{w_i=0}^{1}\sum_{d_i=0}^{1}\begin{pmatrix}1\\ w_i\end{pmatrix}\mathrm{E}[\pi_{d_i|w_i,X_i}\{\mathrm{E}(Y_{id_i}|X_i)-\alpha-\beta w_i\}] = 0.$$

Comparing the solution $\alpha$ and $\beta$ with the true values $\alpha^*$ and $\beta^*$ in model (5.3) reveals that

$$\beta = \beta^* + \mathrm{E}(\pi_{1|1,X}+\pi_{0|0,X}-2)\beta^*, \tag{5.14}$$

$$\alpha = \alpha^* + \mathrm{E}(1-\pi_{0|0,X})\beta^*. \tag{5.15}$$

The asymptotic bias of the naive IPTW estimator of $\beta^*$ thus equals

$$\mathrm{E}(\pi_{1|1,X}+\pi_{0|0,X}-2)\beta^* \tag{5.16}$$

or, when the reclassification probabilities do not depend on $X$, equals $(\pi_{1|1}+\pi_{0|0}-2)\beta^*$. These expressions hold regardless of the choice of index functions $(1\ W_i)$ in (5.4) with $D_i$ replaced by $W_i$. As shown in the Appendix, the DR IPTW estimator has the same asymptotic bias. In contrast, the asymptotic bias of the G-estimator and the OLS estimator additionally depends on the correlation between the exposure variance $\sigma^2_{W|X} \equiv \mathrm{var}(W|X)$ and the reclassification probabilities $\pi_{1|1,X}, \pi_{0|0,X}$. Specifically, the asymptotic bias of the G-estimator of $\beta^*$ is

$$\left\{\frac{\mathrm{cov}(\sigma^2_{W|X},\pi_{1|1,X}+\pi_{0|0,X})}{\mathrm{E}(\sigma^2_{W|X})}+\mathrm{E}(\pi_{1|1,X}+\pi_{0|0,X}-2)\right\}\beta^* \tag{5.17}$$

and the asymptotic bias of the OLS estimator is

$$\left(\frac{\mathrm{cov}\left[\sigma^2_{W|X}+\left\{P(W=1|X)-\mathrm{E}^*(W|X)\right\}^2,\pi_{1|1,X}+\pi_{0|0,X}\right]}{\mathrm{E}\left[\sigma^2_{W|X}+\left\{P(W=1|X)-\mathrm{E}^*(W|X)\right\}^2\right]}+\mathrm{E}(\pi_{1|1,X}+\pi_{0|0,X}-2)\right)\beta^* \tag{5.18}$$

(see the Appendix), where $\mathrm{E}^*(W|X)$ is the fitted value from a linear regression of $W$ on $X$. We conclude that the four considered estimators of the exposure effect of $\beta^*$ are biased when the exposure $D$ is subject to misclassification and the exposure effect differs from 0. Remarkably, they are equally affected by misclassification under linear models if misclassification does not depend on the confounders $X$ used for adjustment. In Section 3.2, we interpret the above bias expressions in more detail and investigate to what extent the bias may differ between the estimators when misclassification

depends on $X$. Finally, note that, for each of the estimators, tests of no exposure effect are valid in the presence of misclassification error (in the sense of preserving the nominal Type I error rate) because the bias expressions (5.16), (5.17) and (5.18) are zero under null hypothesis that $\beta^* = 0$, but that they may be less powerful than in the absence of error.

## 3.2 Simulation study

To investigate the bias of the 4 considered estimators in finite samples, we conducted a simulation experiment. Data were simulated to mimick the birth weight study reported in Hosmer and Lemeshow (2000) with $Y$ representing birth weight, $D$ maternal smoking status during pregnancy and $X$ a confounder score based on maternal age, weight at last menstrual period, race and history of hypertension. Each experiment was based on 1000 replications of random samples of size 189. In each experiment, the covariate $X$ was normally distributed with mean $\mu_X$=3.09 and standard deviation $\sigma_X$=0.32. The observed exposure $W$ and true exposure $D$ were generated to be dichotomous with $P(W = 1|X) = \text{expit}(\theta_0 + \theta_1 X)$ where $\theta_1 = 0.84$ is estimated from the birth weight data and $P(D = 1|W, X) = W\pi_{1|1,X} + (1 - W)(1 - \pi_{0|0,X})$ with $\pi_{1|1,X} = \pi_{0|0,X} = \text{expit}(\eta_0 + \eta_1 X)$. The outcome was chosen to be normally distributed with mean $\beta^* D + X$, $\beta^* = -0.36$ and constant standard deviation 0.65, as estimated from the birth weight data.
Table 5.1 summarizes the results for different choices of the confounder-exposure association, $OR_{W|X} = \exp(\theta_1)$, the exposure mean $\text{E}(W)$, the dependence $\eta_1$ of misclassification on $X$, the average reclassification probability $E(\pi_{1|1,X})$, the correlation $\rho_1$ between $\sigma^2_{W|X}$ and $\pi_{1|1,X}$ and the correlation $\rho_2$ between $\sigma^2_{W|X} + \{\text{E}(W|X) - \text{E}^*(W|X)\}^2$ and $\pi_{1|1,X}$ (see expressions (5.17) and (5.18), respectively). Specifically, we considered the following cases: $\theta_1$ equal to 0.84, 1.68 and $-0.84$, $\text{E}(W)$ equal to 0.40 and 0.6, $\eta_1 = 1/\sigma_X = 3.15$ for the case where $\pi_{1|1,X}$ increases with $X$, $\eta_1 = 0$ when misclassification does not depend on covariates $X$, and $\eta_1 = -1/\sigma_X = -3.15$ for the case where $\pi_{1|1,X}$ decreases with $X$, and finally $E(\pi_{1|1,X})$ equal to 0.80 and 0.90.
As predicted by the theory, all 4 estimators have the same bias when misclassification does not depend on covariates (see the case where $\eta_1 = 0$ in Table 5.1). By noting that $\pi_{1|1,X} = \pi_{0|0,X}$ under our simulations, it follows from expressions (5.14) and (5.17) that the ratio of the bias of the G-estimator versus the (DR) IPTW estimator is

$$\frac{\text{Bias}_G}{\text{Bias}_{(DR)IPTW}} = \frac{\text{cov}(\sigma^2_{W|X}, \pi_{1|1,X})}{\text{E}(\sigma^2_{W|X})\text{E}(\pi_{1|1,X} - 1)} + 1.$$

Therefore, the G-estimator is asymptotically less biased than the (DR) IPTW estimator when $\text{cov}(\sigma^2_{W|X}, \pi_{1|1,X}) \geq 0$ (and thus the negative numerator in the above expression is larger than the negative denominator). This is the case if $\sigma^2_{W|X}$ and $\pi_{1|1,X}$ are both increasing or both decreasing functions of $X$. It follows from the derivative of $\sigma^2_{W|X}$ with respect to $X$,

$$\begin{aligned} \frac{\partial \sigma^2_{W|X}}{\partial X} &= \frac{\partial}{\partial X} P(W=1|X)\{1-P(W=1|X)\} \\ &= \theta_1 \mu_{W|X}(1-\mu_{W|X})(1-2\mu_{W|X}), \end{aligned}$$

that $\sigma^2_{W|X}$ increases with $X$ when the average exposure, $\mu_{W|X} \equiv P(W=1|X)$, is less than 50% and the confounder and exposure have a positive association (i.e. $\theta_1 > 0$). Likewise, $\pi_{1|1,X} = \text{expit}(\eta_0 + \eta_1 X)$ is increasing in $X$ when $\eta_1 > 0$. It follows that the G-estimator is less biased than the (DR) IPTW estimator when the confounder-exposure association is positive (negative), the average exposure is less (more) than 50%, and the reclassification probabilities increase (decrease) with $X$. They are more biased otherwise. Note that these theoretical findings are also reflected in the simulation results. The earlier expressions for the asymptotic bias become intractable when logistic regression models for a dichotomous outcome are considered. We will therefore additionally evaluate the impact of exposure misclassification under logistic regression models by simulation. Each experiment was based on 1000 replications of a random sample in which the observed exposure $W$ and true exposure $D$ were generated as above. The covariate $X$ was standard normally distributed. The outcome $Y$ was generated to be binary with $P(Y=1|D,X) = \text{expit}(\alpha + 2D + X)$. In each simulation experiment, $\alpha$ was chosen such that the outcome mean is 50%, in order to ensure the precision of the estimators to be comparable over the different experiments. The different simulation experiments are characterized by different choices of the confounder-exposure association, $OR_{W|X} = \exp(\theta_1)$, of the exposure mean $\text{E}(W)$ and of the average reclassification probability $E(\pi_{1|1,X}) = E(\pi_{0|0,X})$, with varying dependence $\eta_1$ of misclassification on covariates $X$. Figure 1 shows the relative asymptotic bias of the ordinary maximum likelihood estimator under model (5.10) with $g(X) = \alpha^* + \delta^* X$, the propensity score adjusted estimator (Lunceford and Davidian, 2004), the IPTW estimator and the DR IPTW estimators. Remember that the propensity score adjusted estimator and the IPTW estimator converge to different limiting values because the interpretation of logistic regression parameters changes depending on the covariates included in the model. Specifically, the true causal effect in the marginal structural model (i.e. the causal odds ratio) is equal to $\beta^*_{MSM} = \log \text{odds}(Y_1 = 1) - \log \text{odds}(Y_0 = 1)$ with $P(Y_d = d) = \int P(Y_d = 1|X = x) f_X(x) dx$ with $d = 0, 1$ and $f_X(x)$ the density of $X$, evaluated at $x$. Because the

Table 5.1: *Coverage of 95% confidence intervals for $\beta^*$ and relative bias $E\{(\hat{\beta} - \beta^*)/\beta^*\}$ for 4 considered estimators in function of the confounder-exposure association $OR_{W|X}$, the mean $E(W)$ of the observed exposure, the dependence $\eta_1$ of misclassification on $X$, the average reclassification probability $E(\pi_{1|1,X})$, the correlation $\rho_1$ between $\sigma^2_{W|X}$ and $\pi_{1|1,X}$ (see expression (5.17)) and the correlation $\rho_2$ between $\sigma^2_{W|X} + (P(W = 1|X) - E^*(W|X))^2$ and $\pi_{1|1,X}$ (see expression (5.18)).*

| | | | | | | Coverage | | | | Relative Bias$\times 10^3$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $OR_{W\|X}$ | $E(W)$ | $\eta_1$ | $E(\pi_{1\|1,X})$ | $\rho_1$ | $\rho_2$ | OLS | G | IPTW | DR | OLS | G | IPTW | DR |
| 2.31 | 0.40 | 3.15 | 0.80 | 0.99 | 0.99 | 10.7 | 10.7 | 12.5 | 8.3 | -376 | -376 | -393 | -393 |
| | | | 0.90 | 0.98 | 0.98 | 67.5 | 67.1 | 72.2 | 63.3 | -183 | -183 | -194 | -194 |
| | | 0 | 0.80 | 0.00 | 0.00 | 7.4 | 7.3 | 10.2 | 7.6 | -407 | -407 | -407 | -407 |
| | | | 0.90 | 0.00 | 0.00 | 57.7 | 57.3 | 65.8 | 57.6 | -210 | -210 | -210 | -210 |
| | | -3.15 | 0.80 | -0.77 | -0.78 | 7.9 | 7.7 | 15.0 | 10.5 | -412 | -412 | -399 | -399 |
| | | | 0.90 | -0.68 | -0.67 | 59.9 | 60.1 | 69.5 | 62.5 | -207 | -207 | -200 | -200 |
| | 0.60 | 3.15 | 0.80 | -0.77 | -0.77 | 8.4 | 8.4 | 15.7 | 10.0 | -398 | -398 | -386 | -386 |
| | | | 0.90 | -0.68 | -0.68 | 61.6 | 61.9 | 72.1 | 64.5 | -199 | -199 | -192 | -192 |
| | | 0 | 0.80 | 0.00 | 0.00 | 8.9 | 8.7 | 14.5 | 9.1 | -391 | -391 | -391 | -391 |
| | | | 0.90 | 0.00 | 0.00 | 61.3 | 61.1 | 69.6 | 61.3 | -196 | -196 | -196 | -196 |
| | | -31.5 | 0.80 | 0.99 | 0.99 | 12.0 | 12.1 | 13.7 | 9.7 | -375 | -375 | -392 | -391 |
| | | | 0.90 | 0.98 | 0.98 | 66.4 | 66.1 | 70.7 | 62.2 | -186 | -187 | -197 | -197 |
| 5.34 | 0.40 | 3.15 | 0.80 | 0.93 | 0.93 | 15.2 | 14.6 | 13.5 | 10.1 | -361 | -363 | -395 | -395 |
| | | | 0.90 | 0.93 | 0.93 | 69.6 | 68.7 | 71.5 | 64.4 | -176 | -177 | -199 | -198 |
| | | 0 | 0.80 | 0.00 | 0.00 | 8.4 | 8.7 | 15.0 | 9.7 | -395 | -395 | -395 | -395 |
| | | | 0.90 | 0.00 | 0.00 | 62.3 | 62.4 | 72.2 | 63.0 | -197 | -197 | -197 | -197 |
| | | -3.15 | 0.80 | -0.52 | -0.52 | 8.4 | 8.4 | 16.3 | 10.8 | -413 | -412 | -393 | -393 |
| | | | 0.90 | -0.41 | -0.41 | 58.1 | 57.9 | 72.8 | 61.6 | -210 | -210 | -202 | -202 |
| | 0.60 | 3.15 | 0.80 | -0.53 | -0.53 | 8.0 | 8.2 | 16.4 | 11.6 | -409 | -408 | -390 | -389 |
| | | | 0.90 | -0.42 | -0.42 | 58.7 | 58.7 | 70.3 | 60.8 | -210 | -209 | -201 | -201 |
| | | 0 | 0.80 | 0.00 | 0.00 | 7.8 | 7.4 | 14.1 | 7.8 | -404 | -404 | -404 | -404 |
| | | | 0.90 | 0.00 | 0.00 | 63.0 | 62.6 | 72.4 | 63.6 | -197 | -197 | -197 | -197 |
| | | -3.15 | 0.80 | 0.93 | 0.93 | 16.8 | 16.2 | 16.4 | 11.1 | -357 | -359 | -392 | -391 |
| | | | 0.90 | 0.95 | 0.94 | 70.6 | 69.6 | 71.5 | 64.4 | -173 | -175 | -196 | -196 |
| 0.43 | 0.40 | 3.15 | 0.80 | -0.77 | -0.77 | 7.2 | 7.1 | 12.9 | 8.7 | -412 | -411 | -399 | -399 |
| | | | 0.90 | 0.68 | 0.68 | 59.1 | 59.0 | 70.4 | 62.4 | -205 | -205 | -197 | -197 |
| | | 0 | 0.80 | 0.00 | 0.00 | 8.4 | 8.3 | 12.5 | 9.0 | -396 | -396 | -396 | -396 |
| | | | 0.90 | 0.00 | 0.00 | 60.6 | 60.7 | 67.2 | 60.8 | -201 | -201 | -201 | -201 |
| | | -3.15 | 0.80 | 0.99 | 0.99 | 11.3 | 11.3 | 12.8 | 7.8 | -377 | -377 | -394 | -394 |
| | | | 0.90 | 0.98 | 0.98 | 63.5 | 63.2 | 69.1 | 60.3 | -189 | -189 | -200 | -200 |
| | 0.60 | 3.15 | 0.80 | 0.99 | 0.99 | 10.0 | 9.9 | 12.2 | 7.8 | -383 | -384 | -400 | -400 |
| | | | 0.90 | 0.99 | 0.98 | 64.6 | 64.1 | 69.4 | 61.9 | -186 | -186 | -196 | -196 |
| | | 0 | 0.80 | 0.00 | 0.00 | 7.6 | 7.6 | 11.0 | 7.6 | -403 | -403 | -403 | -403 |
| | | | 0.90 | 0.00 | 0.00 | 64.3 | 64.3 | 72.6 | 64.4 | -191 | -191 | -191 | -191 |
| | | -3.15 | 0.80 | -0.75 | -0.75 | 8.4 | 8.3 | 14.3 | 10.8 | -400 | -400 | -388 | -388 |
| | | | 0.90 | -0.66 | -0.66 | 62.5 | 62.1 | 72.2 | 63.9 | -196 | -196 | -190 | -190 |

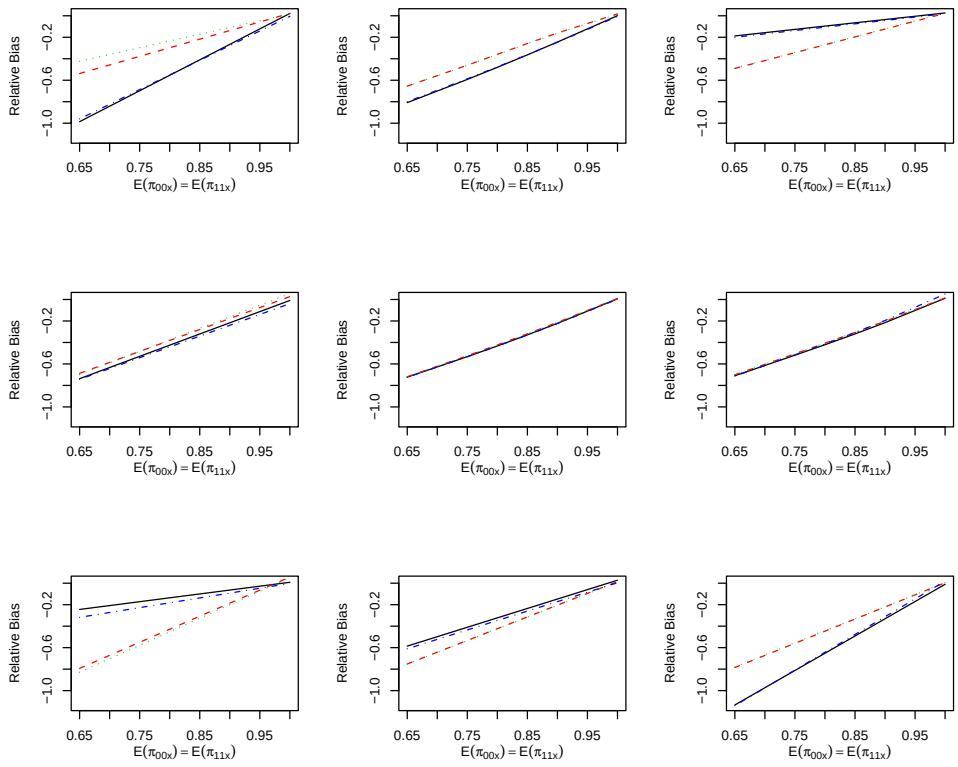



Figure 1: *Relative bias* $E\{(\hat{\beta} - \beta^*)/\beta^*\}$ *as a function of the average reclassification probability* $E(\pi_{1|1,X}) = E(\pi_{0|0,X})$*: Solid line is for the OLS estimator, segmented line is for the IPTW estimator, dotted line is for the DR IPTW estimator and dashed line is for the propensity score adjusted estimator. The confounder-exposure association equals* $OR_{WX} = 4.48$ *in column 1,* $2.72$ *in column 2 and* $0.22$ *in column 3. The dependence* $\eta_1$ *of misclassification on* $X$ *equals* $1$ *in row 1,* $0$ *in row 2 and* $-1$ *in row 3. The correlation* $Cor(\pi_{1|1,X}, X)$ *varies from* $0.98$ *to 0 as* $E(\pi_{1|1,X})$ *changes from* $0.65$ *to* $1.00$ *in row 1, equals zero in row 2, and varies from* $-0.98$ *to 0 as* $E(\pi_{1|1,X})$ *changes from* $0.65$ *to* $1.00$ *in row 3. For all scenarios* $E(W) = 0.80$ *and* $E(Y) = 0.50$*.*

true causal effect estimands differ between the different estimation methods, we will also present the relative bias of the estimators, defined as the ratio of their asymptotic bias to the true causal effect. Figure 1 shows results for 3 cases where $\pi_{1|1,X}$ increases with $X$ (row 1), where misclassification does not depend on covariates (row 2) and where it decreases with $X$ (row 3). It shows the relative bias of the 4 estimators versus the average reclassification probabilities for 9 considered scenarios. Note, in line with the results for linear models, that the propensity score adjusted estimator is more (less) biased than the (DR) IPTW estimator when $\pi_{1|1,X}$ increases (decreases) with $X$, $\mathrm{E}(W) > 0.5$ and the confounder-exposure association is positive (and vice versa when the confounder-exposure association is negative).

# 4 Estimating the causal effect of time-varying exposures

Consider now a longitudinal study whose goal is to estimate the causal effect of a time-varying all-or-nothing exposure $D_{it}$ on outcome $Y_{it}$ at study cycles $t = 0, \ldots, T$ for each of $i = 1, \ldots, n$ independent subjects. For any fixed, non-random treatment history $\overline{d}_t = (d_1, \ldots, d_t)$, let $Y_{t\overline{d}_t}$ be the potential outcome which a subject would have had at time $t$ if, possibly contrary to fact, the exposure history $\overline{D}_t = (D_1, ..., D_t)$ were set to $\overline{d}_t$ through time $t$. For each possible history $\overline{d}_t$ at time $t$, we assume a subject's response $Y_{t\overline{d}_t}$ is well defined and would be observed when $\overline{d}_t = \overline{D}_t$. Suppose again that at each time $t$, the entire confounder history $\overline{X}_t = (X_1, \ldots, X_t)$ for the association between $Y_t$ and $\overline{D}_t$ has been measured, in the sense that for each $s$ and $t$ with $s \leq t$,

$$Y_{t\overline{d}_t} \coprod D_s | \overline{D}_{s-1} = \overline{d}_{s-1}, \overline{X}_s \tag{5.19}$$

for all histories $\overline{d}_t$ (i.e. the no unmeasured confounders assumption). Then we may obtain an estimate for the causal effect of $D_t$ on $Y_t$ at time $t$ by fitting the structural mean model

$$\mathrm{E}\{Y_{t\overline{d}_t} - Y_{t(\overline{d}_{t-1},0)} | \overline{D}_t = \overline{d}_t, \overline{X}_t\} = \beta^* d_t \tag{5.20}$$

for all $t$. This model is similar to the structural nested models of Robins (Robins, 1999), but is less restrictive because it merely specifies the effect of the last exposures and not the effect of earlier exposures. It is therefore mainly of interest in settings where no long-term exposure effects are anticipated. For instance, $D_t$ could indicate condom use at time $t$ and $Y_t$ the resulting HIV status at that time. Additionally, our interest in such models is motivated by the fact that they can be fitted via standard

software. Alternatively, one may develop insight into the causal effect of exposure on outcome by fitting a marginal structural model (Robins, 1999 and 1986); e.g.

$$\mathrm{E}(Y_{t\overline{d}_t}) = \alpha_t^* + \beta^* d_t. \tag{5.21}$$

Note that model (5.21) postulates only the last exposure $d_t$ to have an effect on the outcome at time $t$. In that sense, this model is more restrictive than model (5.20), which makes no assumptions about the effect of early exposures. It is thus not surprising that model (5.20) with the additional restriction that there is not effect of early exposures; that is

$$\mathrm{E}\{Y_{t\overline{d}_s} - Y_{t(\overline{d}_{s-1},0)} | \overline{D}_s = \overline{d}_s, \overline{X}_s\} = 0$$

for all $s < t$, implies model (5.21). In contrast, when early exposures themselves affect the outcome, on must extend model (5.21) by explicitly parameterizing these effects; for instance, by fitting model

$$\mathrm{E}(Y_{t\overline{d}_t}) = \alpha_t^* + \beta_1^* d_t + \beta_2^* \sum_{s=1}^{t-1} d_s. \tag{5.22}$$

Standard regression approaches for longitudinal data are unable to adjust for time-varying confounders, and thus unable to fit models (5.21) and (5.22), even when all the relevant time-varying confounders have been measured (Robins, 1999 and 1986; Hernán et al., 2002). This is because standard regression approaches ignore that the exposure and time-varying confounders may mutually influence each other over time. Specifically, exposure and outcome in longitudinal studies may be influenced by previous exposures and outcomes and possibly also by other time-varying confounders which may be intermediate on the causal path from the exposure to the outcome (Robins, 1999 and 1986). In this section, we briefly review IPTW estimation of the causal parameters indexing the marginal structural model (5.21) and likewise model model (5.22). In addition, we introduce G-estimators for the causal parameters indexing the structural mean model (5.20), which can be obtained with standard software. In Section 5, we then examine the impact of exposure misclassification on these estimators.

## 4.1 IPTW estimator

Estimates for $\beta^*$ in model (5.21) may be obtained by extending the IPTW approach of Section 2.1 to time-varying exposures, as discussed in (Robins et al., 2000;

Robins, 1999). Specifically, they may be obtained by fitting the sequential conditional mean model

$$\mathrm{E}(Y_t|\overline{D}_t) = \alpha_t^* + \beta^* D_t$$

for $t = 1, \ldots, T$ using generalized estimating equations (GEE) with independence working correlation (see Vansteelandt, 2007, for more efficient estimation approaches), where the contribution of each subject $i$'s data at time $t$ is weighted by

$$SW_{it} = \prod_{s=1}^{t} \frac{P(D_{is}|\overline{D}_{i,s-1})}{P(D_{is}|\overline{D}_{i,s-1}, \overline{X}_{is})}. \tag{5.23}$$

Assuming that $\alpha_t^* = \alpha^{(0)*} + \alpha^{(1)*}t$ for illustration, IPTW estimation thus proceeds by solving the following estimating equation for $\beta$

$$\sum_{i=1}^{n} U_{I_{it}}(\alpha^{(0)}, \alpha^{(1)}, \beta) = \sum_{i=1}^{n}\sum_{t=1}^{T} \begin{pmatrix} 1 \\ t \\ D_{it} \end{pmatrix} SW_{it} \left(Y_{it} - \alpha^{(0)} - \alpha^{(1)}t - \beta D_{it}\right) = 0. \tag{5.24}$$

In (5.23), the denominator represents the probability that subject $i$ follows his/her own exposure history, given the history of time-varying confounders. As the denominator probabilities are unknown, they may be estimated by fitting logistic regression models for the observed exposure $D_t$ at each time $t$, given $\overline{D}_{t-1}$ and $\overline{X}_t$. When the exposure regime is monotone (i.e., when $D_t = 0$ implies $D_s = 0$ for all $s > t$), this may happen by fitting a logistic regression model of the form

$$P(D_{it} = 1|D_{it-1} = 1, \overline{X}_{it}; \theta) = \mathrm{expit}\left(\theta_0 + \theta_1 t + \theta_2 X_{it}\right) \tag{5.25}$$

for $t = 1, ..., T$. Likewise, the numerator represents the unconditional probability that subject $i$ follows his/her own exposure history and (for monotone regimes) may be estimated by fitting a logistic regression model of the form

$$P(D_{it} = 1|D_{it-1} = 1; \phi) = \mathrm{expit}\left(\phi_0 + \phi_1 t\right) \tag{5.26}$$

for $t = 1, ..., T$. The denominator weights ensure that after inverse weighting, there is no association between $D_s$ and $\overline{X}_s$ for $s = 1, ..., t$ at each time $t$, so that there are no further time-varying confounders. Correct specification of the regression models corresponding to these weights is henceforth important. The numerator weights are used to improve the finite-sample performance of the IPTW estimator, which may be poor when the denominator probabilities are small for some subjects. Misspecification of these numerator weights does not affect the validity of the approach.

## 4.2 G-estimator

Estimates for the causal parameters $\beta^*$ in model (5.20) may be obtained by repeating the two-step G-estimation procedure of Section 2.2 at each time point in the study. Specifically, model (5.20) implies

$$\mathrm{E}\{Y_t - \beta^* D_t | \overline{D}_t, \overline{X}_t\} = \mathrm{E}\{Y_{t(\overline{D}_{t-1},0)} | \overline{D}_t, \overline{X}_t\} = \mathrm{E}\{Y_{t(\overline{D}_{t-1},0)} | \overline{D}_{t-1}, \overline{X}_t\} \qquad (5.27)$$

where the last equality follows from the no unmeasured confounders assumption (5.19). It then follows from (5.27) that

$$\mathrm{E}\left[\{D_t - \mathrm{E}(D_t|\overline{D}_{t-1}, \overline{X}_t)\}\{Y_t - \beta^*(D_t - \mathrm{E}(D_t|\overline{D}_{t-1}, \overline{X}_t)) - g_t(\overline{D}_{t-1}, \overline{X}_t)\}\right] = 0$$

for each function $g_t(\overline{D}_{t-1}, \overline{X}_t)$ of the past exposure history $\overline{D}_{t-1}$ and covariate history $\overline{X}_t$. Summing this over all time points yields an unbiased estimating equation for $\beta^*$:

$$\begin{aligned} 0 = \sum_{i=1}^{n} U_{G_i}(\beta) \quad = \quad & \sum_{i=1}^{n}\sum_{t=1}^{T} \{D_{it} - \mathrm{E}(D_{it}|\overline{D}_{i,t-1}, \overline{X}_{it})\} \\ & \times [Y_{it} - \beta(D_{it} - \mathrm{E}(D_{it}|\overline{D}_{i,t-1}, \overline{X}_{it})) - \\ & g_t(\overline{D}_{i,t-1}, \overline{X}_{it})] \end{aligned} \qquad (5.28)$$

where $D_0 = 0$. Choosing

$$g_t(\overline{D}_{i,t-1}, \overline{X}_{it}) = \mathrm{E}(Y_{it}|\overline{D}_{i,t-1}, \overline{X}_{it})$$

yields an estimating equation which is unbiased when at each time $t$, either the working model for $\mathrm{E}(D_{it}|\overline{D}_{i,t-1}, \overline{X}_{it})$ is correctly specified, or the working model for $\mathrm{E}(Y_{it}|\overline{D}_{i,t-1}, \overline{X}_{it})$. Estimators which are consistent and asymptotically normally distributed when at each time $t = 1, ..., T$ one of these 2 working models is correctly specified, are referred to as $2^T$ multiply-robust estimators (Vansteelandt et al., 2007). Solving (5.29) with $g_t(\overline{D}_{i,t-1}, \overline{X}_{it}) = \gamma_0^* + \gamma_1^* t + \gamma_2^* X_{it}$ for unknown parameters $\gamma_0^*, \gamma_1^*, \gamma_2^*$ is equivalent to fitting the sequential conditional mean model

$$\mathrm{E}\{Y_{it}|\overline{D}_{it}, \overline{X}_{it}; \gamma_0, \gamma_1, \gamma_2, \beta\} = \gamma_0 + \gamma_1 t + \gamma_2 X_{it} + \beta\{D_{it} - \mathrm{E}(D_{it}|\overline{D}_{it-1}, \overline{X}_{it})\}. \quad (5.29)$$

The parameters in this model can be estimated using standard software by solving independence generalized estimating equations (GEE) (see Vansteelandt, 2007, for more efficient estimation approaches).

# 5 The impact of time-varying exposure misclassification

In longitudinal repeated measures data, reclassification probabilities of time-varying exposures at each time $t$ should be formulated in function of the true and observed exposure history and in function of the history of time-varying confounders. That is, at each time $t$, we must postulate the reclassification probabilities as

$$P(D_t = d_t | W_t = w_t, \overline{D}_{t-1}, \overline{W}_{t-1}, \overline{X}_t).$$

In practice, because little information is typically available about these probabilities, we will assume for simplicity that time-varying reclassification probabilities are independent of covariates. In addition, as is common in analyses of time-varying exposures (Dawson and Lavori, 2002), we will consider monotone exposure regimes. Further assuming that misclassification does not depend on the history of true or observed exposure, we thus find that the reclassification probabilities at each time $t$ equal

$$P(D_t = d_t | W_t = w_t, \overline{D}_{t-1}, \overline{W}_{t-1}, \overline{X}_t) = \begin{cases} \pi_{d_t|w_t} & \text{if} \quad \overline{D}_{t-1} = 1 \\ I(d_t = 0) & \text{if} \quad \overline{D}_{t-1} = 0 \end{cases} \tag{5.30}$$

where $I(d_t = 0)$ is 1 if $d_t = 0$ and 0 otherwise. Specifically, reclassification probabilities can be expressed $P(D_t = 1 | \overline{D}_{t-1} = 1, \overline{W}_t = 1, \overline{X}_t) = \pi_{1|1}$ and $P(D_t = 0 | W_t = 0, \overline{D}_{t-1} = 1, \overline{W}_{t-1}, \overline{X}_t) = \pi_{0|0}$. It also follows that $P(D_t = 0 | \overline{W}_t, \overline{D}_{t-1} = 0, \overline{X}_t) = 1$.

As in the previous section, we will assume that misclassification is non-differential in the sense that $Y_{t\overline{d}_t} \coprod W_t | \overline{D}_t, \overline{W}_{t-1}, \overline{X}_t$ for each $t = 1, \ldots, T$ and for all possible exposure histories $\overline{d}_t$. This assumption states that the degree of misclassification is independent of the outcome at each time $t$.

To obtain the asymptotic bias of the IPTW estimator of Section 4.1 and the G-estimator of Section 4.2 with $D_t$ replaced by $W_t$, we calculated the limiting value of $\beta$ that solves the corresponding expected estimating equation. The latter involves the conditional distribution of $X_t$ given $\overline{D}_{t-1}, \overline{X}_{t-1}$ and $X_t$ given $\overline{X}_{t-1}$ in a complex way, which is difficult to estimate well in practice because of its possible high dimensionality. We will therefore report simulation studies to investigate the asymptotic bias of both estimators.

In each of 1000 simulation runs, time-varying covariates $X_t$ were generated as $X_t = \rho X_{t-1} I(t > 1) + D_{t-1} I(t > 1) + N(0, \sigma^2) I(t = 1)$ where the autocorrelation $\rho$ was chosen to equal $(\sigma^2 - 1)/\sigma^2$, such that the variance of $X_t$ is constant at each time $t$. The time-varying observed exposures $W_t$ and true exposures $D_t$ were generated to

be binary with $P(W_t = 1|\overline{X}_t, \overline{W}_{t-1}, \overline{D}_{t-1}) = \text{expit}(\theta_0 + 0.03t + \theta_2 X_t)$ and $P(D_t = 1|\overline{X}_t, \overline{W}_t, \overline{D}_{t-1}) = W_t\pi_{11} + (1-W_t)(1-\pi_{00})$. Time-varying outcomes were generated as $Y_t = -1.5 + 1.5t + \gamma X_t + \beta^* D_t + N(0,1)$ with $\beta^* = 2$. Note that both the no unmeasured confounders asumption (5.19) and the non-differential misclassification assumption hold under our simulations.

In all simulations, we fitted the following model

$$\text{E}(Y_{\overline{td}_t}) = \alpha_0 + \alpha_1 t + \sum_{k=1}^{t-1} \alpha_{k+1} d_k + \beta d_t \tag{5.31}$$

for the IPTW estimator by using the corresponding linear regression model with stabilized weights. It can be shown that this model is correctly specified under the data-generating mechanism of our simulation experiment, with $\beta = 2$. When fitting this model, we used a stepwise model building approach to select appropriate past exposure histories to be included in the model at each time $t$. Further, the following correctly-specified model

$$\text{E}(Y_t|\overline{D}_t, \overline{X}_t) = \alpha_0 + \alpha_1 t + \gamma X_t + \beta(D_t - \text{E}(D_t|\overline{D}_{t-1}, \overline{X}_t)) \tag{5.32}$$

was fitted using ordinary generalized estimating equations (GEE) (Liang and Zeger, 1986) with independence working correlation to obtain the G-estimator. We have carried out different simulations by various choices of $(\theta_2, \gamma, \rho)$. Here, $\theta_2$ measures the strength of the association between confounders and exposures, with $OR_{W_t|X_t} = \exp(\theta_2)$. Further, $\gamma$ measures the strength of the association between confounders and outcome, and $\rho$ is the autocorrelation of $X_t$ at each time $t = 1, ..., 10$. Simulation studies are repeated for 6 choices of the parameter values, listed in Table 5.2. In each scenario, $\theta_0$ was selected to keep the mean value of $W_t$ approximately fixed over the different time points $t = 1, ..., 10$ (see Table 5.2). Figure 2 shows the relative bias $\text{E}\{(\hat{\beta}_t - \beta^*)/\beta^*\}$ of the IPTW estimator and the G-estimator versus the end-of-study time $t$ when the association between outcomes and covariates is $\gamma = 1$. The figure considers 3 cases with reclassification probabilities $\pi_{0|0} = \pi_{1|1}$ equal to 1, 0.9 and 0.8, as well as various cases of confounder-exposure association and correlation between $X_t$ and $X_{t+1}$ at each time $t$. Note that we also show results for the unadjusted estimator (i.e. the IPTW estimator with weights set to 1) so as to illustrate the degree of confounding bias in each simulation. Figure 2 shows that the IPTW estimator and the G-estimator are equally affected by misclassification error when the confounder-exposure association $\theta_2$ is equal to 0.5 and 1 (left 4 panels). When this association becomes severe ($\theta_2 = 2$), the G-estimator becomes significantly less biased than the IPTW estimator. This is because the weights in the IPTW estimator are then more

variable so that the IPTW estimator has greater finite sample bias, even in the absence of measurement error. This increased variability of the weights is displayed in Table 5.2, where we show the empirical ratio $P_{0.975}/P_{0.025}$ of the 97.5th and 2.5th percentile of the distribution of the stabilized weights $SW(t)$ over all time points $t = 1, ..., 10$. We also found the discrepancies between the IPTW-estimator and the G-estimator to be slightly increased with higher autocorrelation. This is because this higher autocorrelation leads to less information and thus an IPTW estimator which is slightly more biased in finite samples. Finally, Figure 3 shows the relative standard deviation $SD(\hat{\beta}_t)/\beta^*$ of the different estimators and suggests the increased efficiency of the G-estimator over the IPTW estimator at each time.

Table 5.2: *Simulation set-up: parameters, ratio of the 97.5th and 2.5th percentile of the stabilized weight distribution and mean value of* $W_t$ *at time* $t = 1$ *and* $t = 10$.

| $(\theta_2, \gamma, \rho)$ | $(\frac{P_{1,0.975}}{P_{1,0.025}}, \frac{P_{10,0.975}}{P_{10,0.025}})$ | $(\mathrm{E}(W_1), \mathrm{E}(W_{10}))$ |
|---|---|---|
| (0.5,1.0,0.3) | (3.05,6.35) | (0.17,0.79) |
| (1.0,1.0,0.3) | (7.08,34.28) | (0.14,0.75) |
| (2.0,1.0,0.3) | (13.39,76.77) | (0.10,0.67) |
| (0.5,1.0,0.6) | (2.80,5.78) | (0.15,0.76) |
| (1.0,1.0,0.6) | (8.44,35.06) | (0.16,0.70) |
| (2.0,1.0,0.6) | (12.33,144.22) | (0.13,0.62) |

# 6 Discussion

In this paper, we have investigated the impact of exposure misclassification on the asymptotic bias of different effect estimators when exposure effects are confounded by measured covariates. Our interest in this stems from the fact that we anticipated the different estimators to be differently affected by misclassification error because some of them not only use the exposure as a covariate in a regression model, but also rely on estimates of the expected exposure, given background covariates. However, in contrast to our anticipation, we found OLS estimators, (DR) IPTW estimators and G-estimators to be equally affected by misclassification error under linear models (and approximately under logistic models) when misclassification is unrelated to the considered confounders. Differences in asymptotic bias between the different estimators arise when misclassification is related to the confounders. Indeed, now the bias of the OLS estimator and G-estimator becomes dependent upon the degree of

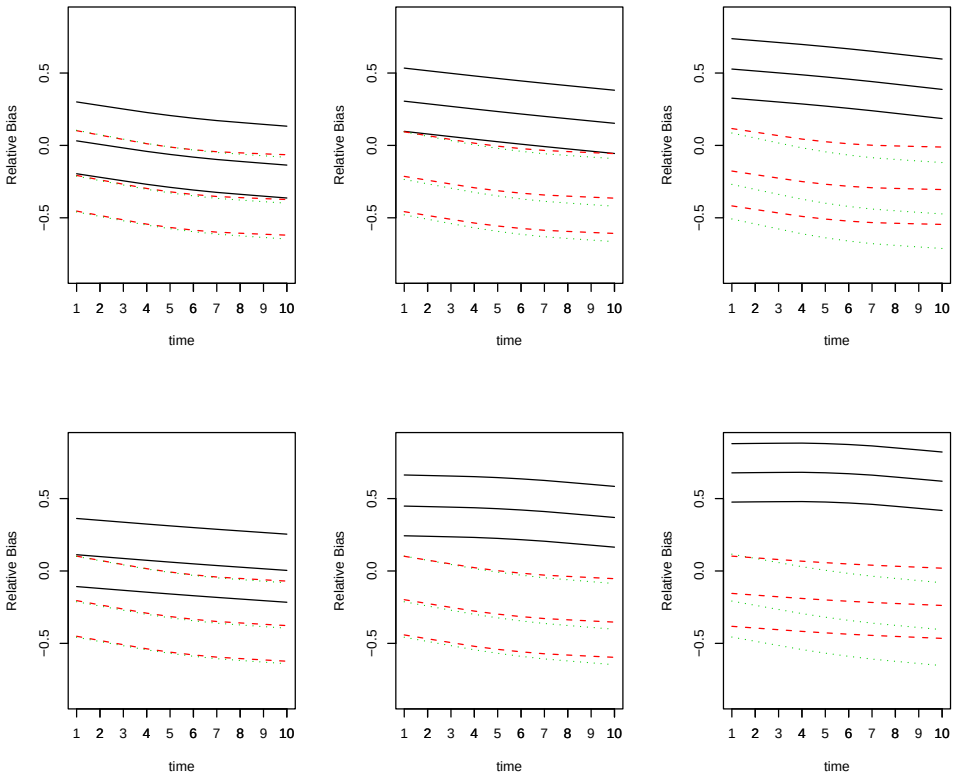



Figure 2: *Relative bias* $E\{(\hat{\beta}_t - \beta^*)/\beta^*\}$ *as a function of end-of-study time* $t$*, for the unadjusted estimator (solid line), the IPTW estimator with stabilized weights (segmented line) and the G-estimator (dotted line), when the reclassification probabilities are* $\pi_{0|0,X} = \pi_{1|1,X} = 0.8, 0.9, 1$ *and* $\gamma$ *takes the value* $1$*. For the IPTW estimator and the G-estimator from top to bottom of each plot, the lines correspond to* $\pi_{0|0,X} = \pi_{1|1,X} = 1, 0.9$ *and* $0.8$*. In rows 1 and 2 from left to right,* $\theta_2$ *takes the values* $0.5, 1$ *and* $2$ *respectively. The autocorrelation* $\rho$ *equals* $0.3$ *in row 1 and* $0.6$ *in row 2.*

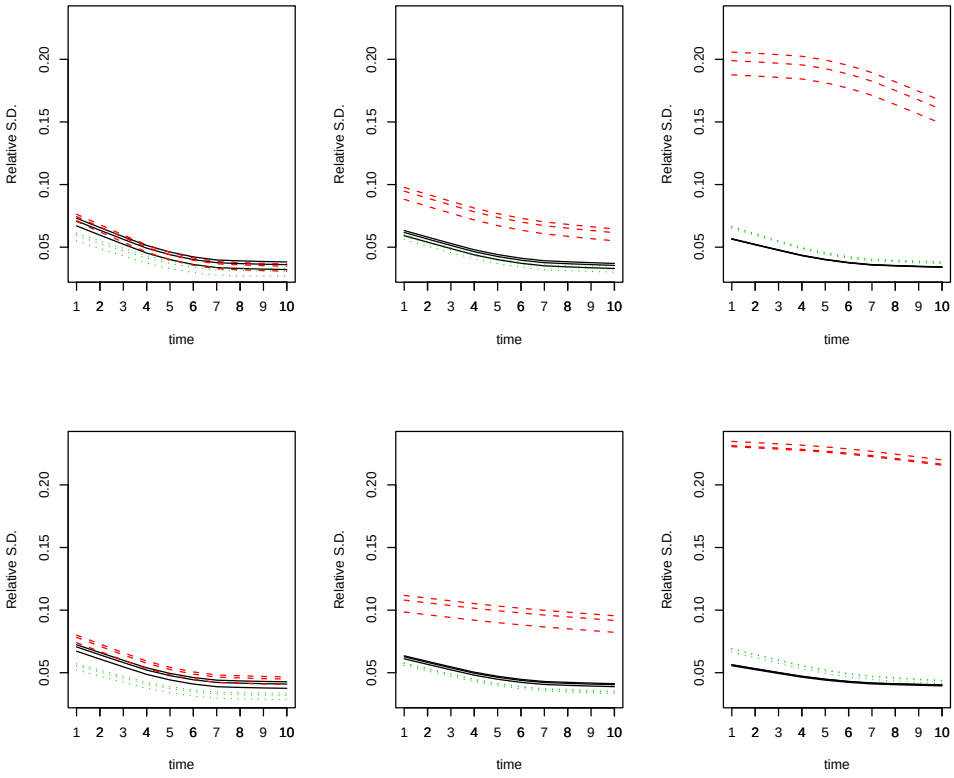



Figure 3: *Relative standard deviation* $SD(\hat{\beta}_t)/\beta^*$ *as a function of end-of-study time* $t$*, for the unadjusted estimator (solid line), the IPTW estimator with stabilized weights (segmented line) and the G-estimator (dotted line), when reclassification probabilities are* $\pi_{0|0,X} = \pi_{1|1,X} = 0.8, 0.9, 1$ *and* $\gamma$ *takes the value* $1$*. In rows 1 and 2 from left to right* $\theta_2$ *takes the values* $0.5, 1$ *and* $2$ *respectively. The autocorrelation* $\rho$ *equals* $0.3$ *in row 1 and* $0.6$ *in row 2.*

confounders-exposure association, unlike the bias of the (DR) IPTW estimator. The latter may be due to the inherent impact of inverse weighting, which is to eliminate the association between exposure and confounder.

# Appendix 5.A: Derivation of asymptotic bias expressions for the considered causal effect estimators

In this Appendix, we calculate the asymptotic bias of the four considered estimators by solving the expected estimating function of these estimators with $D$ replaced by $W$. For IPTW estimators, we first take the expectation of the estimating function (with general index function $h(w) = (h_1(w)\ h_2(w))$ over the conditional distribution of $D$, given $(Y_d, W, X)$ and the conditional distribution of $W$, given $(Y_d, X)$. Upon noting that $Y_d \perp\!\!\!\perp W|D, X$ for $d = 0, 1$ together with the assumption of no unmeasured confounders (5.1) implies that $Y_d \perp\!\!\!\perp W|X$ for $d = 0, 1$, this yields,

$$\begin{aligned}
0 = \mathrm{E}(U_I(\beta^*)) &= \mathrm{E}\left[\sum_{d=0}^{1} h(W)\frac{P\left(D=d|W,Y_d,X\right)}{P(W|X)}\left(Y_d-\alpha-\beta W\right)\right] \\
&= \mathrm{E}\left[\sum_{w=0}^{1}\sum_{d=0}^{1} h(w)P(D=d|W=w,X)(Y_d-\alpha-\beta w)\right] \\
&= \mathrm{E}\left[\sum_{w=0}^{1}\sum_{d=0}^{1} h(w)\pi_{d|w,X}\left\{g(X)+\beta^* d-\alpha-\beta w\right\}\right] \\
&= \sum_{w=0}^{1}\sum_{d=0}^{1} h(w)\mathrm{E}\{\pi_{d|w,X}\left(g(X)-\alpha\right)\} + \mathrm{E}(\pi_{d|w,X})(\beta^* d-\beta w)\big).
\end{aligned}$$

Solving the resulting set of equations

$$(h_j(0)+h_j(1))(\alpha^*-\alpha)+[h_j(0)(1-\mathrm{E}(\pi_{0|0,X}))+h_j(1)\mathrm{E}(\pi_{1|1,X})]\beta^*-h_j(1)\beta=0$$

for $j=1,2$ yields (5.14) and (5.15), regardless of the choice of index functions.

The estimating function for the doubly robust (DR) IPTW estimator is

$$\begin{aligned} U_D(\alpha,\beta) &= \frac{h(W)(Y-\alpha-\beta W)}{P(W|X)}-\mathrm{E}\left\{\frac{h(W)(Y-\alpha-\beta W)}{P(W|X)}|W,X\right\} \\ &\quad +\mathrm{E}\left\{\frac{h(W)(Y-\alpha-\beta W)}{P(W|X)}|X\right\} \end{aligned}$$

where the last 2 terms have mean zero when, as we assume, the model for $P(W|X)$ is correctly specified. It follows that $\mathrm{E}\{U_D(\alpha,\beta)\}=\mathrm{E}\{U_I(\alpha,\beta)\}$ and thus that DR IPTW and IPTW estimators have the same bias.

To find the asymptotic bias of G-estimator, we take the expectation of the estimating function (5.11) (with $D$ replaced by $W$) over the conditional distribution $Y$, given $(D,W,X)$ and then over the conditional distribution $D$, given $(W,X)$:

$$\begin{aligned} 0 &= \mathrm{E}\{U_G(\beta)\}=\mathrm{E}\left[\{W-\mathrm{E}(W|X)\}\{\mathrm{E}(Y|D,W,X)-\beta W-\mathrm{E}(Y-\beta W|X)\}\right] \\ &= \mathrm{E}\left[\{W-\mathrm{E}(W|X)\}\{\beta^*(D-\mathrm{E}(D|X))-\beta(W-\mathrm{E}(W|X))\}\right] \\ &= \mathrm{E}\left[\{W-\mathrm{E}(W|X)\}^2\{\beta^*(\pi_{1|1,X}-\pi_{1|0,X})-\beta\}\right] \\ &= \mathrm{E}\left[\beta^*\sigma^2_{W|X}(\pi_{1|1,X}-\pi_{1|0,X})-\beta\sigma^2_{W|X}\right]. \end{aligned}$$

Solving this equation yields

$$\begin{aligned} \beta &= \frac{\beta^*}{\mathrm{E}\{\sigma^2_{W|X}\}}\mathrm{E}\{\sigma^2_{W|X}(\pi_{1|1,X}+\pi_{0|0,X}-1)\} \\ &= \beta^*\left[\frac{Cov(\sigma^2_{W|X},\pi_{1|1,X}+\pi_{0|0,X}-1)}{\mathrm{E}(\sigma^2_{W|X})}+\mathrm{E}(\pi_{1|1,X}+\pi_{0|0,X}-1)\right]. \end{aligned}$$

From Robins et al.(1992, page 486) the OLS estimator can be obtained by replacing, $\mathrm{E}(W|X)$ with $\mathrm{E}^*(W|X)$, the fitted value from a linear regression of $W$ on $X$ in the estimating equation (5.11). It then follows that $\sigma^2_{W|X}$ in the above expression should be replaced with

$$\begin{aligned} \mathrm{E}\{(W-\mathrm{E}^*(W|X))^2|X\} &= \mathrm{E}\{(W-\mathrm{E}(W|X)+\mathrm{E}(W|X)-\mathrm{E}^*(W|X))^2|X\} \\ &= \sigma^2_{W|X}+\{\mathrm{E}(W|X)-\mathrm{E}^*(W|X)\}^2. \end{aligned}$$

# Appendix 5.B: T ime-varying exposures

Here, we show that when there is no interaction effect between $D_t$ and $X_t$ at each time $t$, model (5.20) with the additional restriction that there is no effect of early exposures, that is, $\mathrm{E}\{Y_{t(d_{t-1},0)} - Y_{t0}|\overline{D}_{t-1}, \overline{X}_{t-1}\} = 0$, implies model (5.21). It follows from model (5.20) that,

$$\mathrm{E}\{Y_{t\overline{D}_t}|\overline{D}_t, \overline{X}_t\} = \mathrm{E}\{Y_{t(\overline{D}_{t-1},0)}|\overline{D}_{t-1}, \overline{X}_t\} + \beta^* D_t$$

by no unmeasured confounders assumption, we will have

$$\mathrm{E}\{Y_{td_t\overline{D}_{t-1}}|\overline{D}_{t-1}, \overline{X}_t\} = \mathrm{E}\{Y_{t(d_{t-1}\overline{D}_{t-2},0)}|\overline{D}_{t-2}, \overline{X}_t)\} + \beta^* d_t.$$

By taking expectation on the latter expression with respect to $X_t$,

$$\mathrm{E}\{Y_{td_td_{t-1}\overline{D}_{t-2}}|\overline{D}_{t-2}, \overline{X}_{t-1}\} = \mathrm{E}\{Y_{t(d_{t-1}d_{t-2}\overline{D}_{t-3},0)}|\overline{D}_{t-3}, \overline{X}_{t-1})\} + \beta^* d_t$$

by repeating this process we will obtain $\mathrm{E}\{Y_{t\bar{d}_t}\} = \mathrm{E}\{Y_{t(\bar{d}_{t-1},0)}\} + \beta^* d_t$. Therefore by supposing that $\mathrm{E}\{Y_{t(\bar{d}_{t-1},0)}\} = \alpha_t$ implies model (5.21).
Now we show that the expected estimating equations of both estimators involve the conditional distribution $X_t$ given $\overline{D}_{t-1}, \overline{X}_{t-1}$ and $X_t$ given $\overline{X}_{t-1}$. For illustration $t = 2$, model (5.21) can be written $\mathrm{E}\{Y_{2\bar{d}_2}\} = \alpha_2^* + \beta^* d_2$. The estimating equation of (5.24) when $W_t$ is observed instead of $D_t$ with non stabilized weights can be obtained

$$U_{I2}(\beta) = \begin{pmatrix} 1 \\ 1 \\ W1 \end{pmatrix} \frac{1}{P(W1|X_1)}(Y_{1d_1} - \alpha_1 - \beta W_1) + \begin{pmatrix} 1 \\ 2 \\ W2 \end{pmatrix} \frac{1}{P(W1|X_1)P(W2|W_1, \overline{X}_2)}(Y_{2d_2} - \alpha_2 - \beta W_2).$$

By non-misclassification assumption, the expected estimating equation of $U_{I2}(\beta)$ follows that,

$$0 = \mathrm{E}(U_I 2(\beta)) = \mathrm{E}\{\sum_{d=0}^{1} \begin{pmatrix} 1 \\ 1 \\ W1 \end{pmatrix} \frac{P(D_1 = d1|W_1, X_1)}{P(W1|X_1} (Y_{1d_1} - \alpha^{(1)} - \beta W_1) + \sum_{d_1=0}^{1}\sum_{d_2=0}^{1} \begin{pmatrix} 1 \\ 2 \\ W2 \end{pmatrix} \frac{I(D_1 = d_1)P(D_2 = d_2|\overline{W}_2, D_1 = d_1, \overline{X}_2)}{P(W_1|X_1)p(W_2|W_1, \overline{X}_2)} (Y_{2\bar{d}_2} - \alpha^{(2)} - \beta W_2)\}$$

and the limiting values $(\alpha_1, \alpha_2, \beta)$ that solve the resulting expected estimating equation should be obtained by solving equation below which is involved the conditional distribution of $X_2$ given $D_1$, $W_1$,$X_1$ and $X_2$ given $X_1$,

$$0 = \mathrm{E}(U_{I2}(\beta)) = \mathrm{E}\{\sum_{d=0}^{1}\sum_{w_1=0}^{1} \begin{pmatrix} 1 \\ 1 \\ w1 \end{pmatrix} \pi_{d_1|w_1}(Y_{1d_1} - \alpha_0 - \alpha_1 - \beta w_1) + \sum_{d_1=0}^{1}\sum_{w_1=0}^{1}\sum_{w2_1=0}^{1}\sum_{d_2=0}^{1} \begin{pmatrix} 1 \\ 2 \\ w2 \end{pmatrix} \pi_{d_1|w_1}\pi_{d_2|w_2} \frac{P(X_2|D_1 = d_1, W_1 = w_1, X_1)}{P(X_2|X_1)} \times (Y_{2\bar{d}_2} - \alpha_0 - 2\alpha_1 - \beta w_2)\}.$$

The limiting value $\beta$ that solves the resulting expected limiting estimating equation (5.29) of G-estimation is also involved the conditional distribution $X_2$ given $D_1$, $W_1$ and $X_1$ and $X_2$ given $X_1$.

# Chapter 6

# Discussion

## Summary

This chapter gives a final discussion and plans for future work. We review the results which were presented in the previous chapters. We will explain how the methods discussed in this thesis can be extended and what obstacles that one might be confronted with. Moreover, we will discuss plans for future work.

## Introduction

This thesis has addressed two problems: the problem of measurement error in exposures and of measured or unmeasured confounding of the effect of exposure on outcome. As stated, a common strategy for dealing with two these problems is to use instrumental variable estimation methods.
In this thesis, we have first given an expository review of IV-estimators for the causal odds ratio. Specifically, we have focused on exact estimators as well as a number of popular approximate ones, without being exhaustive; in particular, we have omitted estimators based on principal stratification (e.g. Abadie, 2003; Ten Have et al., 2003) as this approach does not allow a flexible treatment of continuous exposures and is rather artificial in the context of Mendelian randomization studies (Didelez, Meng and Sheehan, 2008). Our results have shown that the concerns of Robins and Rotnitzky (2004) about incongeniality of the model of Vansteelandt and Goetghebeur (2003) can be overruled by leaving the main effect of the IV in their association model unrestricted. For general IVs, leaving its main effect unrestricted requires the use of

generalized additive association models. The performance of the resulting estimator remains to be studied. Our simulation studies complemented recent studies by Didelez, Meng and Sheehan (2008), Palmer et al. (2008) and Rassen et al. (2008), but include results on the bias and efficiency of exact IV-estimators. They have revealed that these 'exact' estimators tend to outperform 'approximate' estimators of the causal odds ratio that are commonly used in the literature on Mendelian randomization. Of all considered estimators, the Exact IV-estimators are the only ones which are asymptotically unbiased, although they may have an important finite-sample bias when there is limited information (e.g. due to low prevalence, weak IV, small sample size, ...). The Approximate IV-estimator tended to outperform standard logistic regression when there was confounding of a sufficient magnitude, and was doing especially well at the causal null hypothesis. This estimator has a number of attractions over the Exact IV-estimators in that it can be used for the analysis of case-control data (Smith et al., 2005) and lends itself particularly well to meta-analyses based on summary measures (Minelli et al., 2004). In addition, it can be extended to the analysis of time-varying exposures. Indeed, consider the following logistic structural nested mean model

$$\text{logit}E(Y_{t(\overline{x}_s 0)}|\overline{X}_s = \overline{x}_s, R) - \text{logit}E(Y_{t(\overline{x}_{s-1} 0)}|\overline{X}_s = \overline{x}_s, R) = \psi^* I(s=t)x_t + \gamma^* I(s<t)x_s$$

where $Y_{t(\overline{x}_s 0)}$ denotes the counterfactual outcome that would be observed at time $t$ if, possibly contrary to fact, the exposure history $\overline{X}_t = (X_1, ..., X_t)$ equaled $(\overline{x}_s, 0, ..., 0)$, and $R$ is instrumental variable (IV) the effect of $X_t$ on $Y_t$ at time $t$, which by definition, satisfies the following properties (a) $R$ is associated with $X_t$; (b) $R$ affects the outcome $Y_t$ only through $X_t$ (i.e. often referred to as the exclusion restriction); (c) the association between $R$ and $Y_t$ is unconfounded. This model allows for the short term effect $\psi^*$ to differ from the long term effect $\gamma^*$. Using similar approximations in one time point, it can be shown that this model implies

$$\text{logit}E(Y_t|R) = \omega_t^* + \psi^* E(X_t|R) + \gamma^* \sum_{s=1}^{t-1} E(X_s|R)$$

where $Y_t$ is the observed outcome at time $t > 0$. Having obtained estimates of $E(X_s|R)$ for $s > 0$, this model can be fitted, and thus estimates of $\psi^*$ and $\gamma^*$ can be obtained using standard software for marginal models. The resulting estimators continue to share the local robustness property of being consistent at the causal null hypothesis. It remains to be evaluated to what extent the approximation errors propagate with time and thus how prone to bias these estimators are away from the causal null hypothesis. It additionally remains to be studied whether similar adjustments as for the Adjusted IV-Estimator (possibly including the previously suggested

corrections for measurement error) can remedy some of the bias of the resulting estimator. When the exposure and error are additive, then the Adjusted IV-estimator is still prone to some bias because the model for $E(X|R)$ may be misspecified and because, even when it is correctly specified, $E(X|R)$ is not known in practice so that $X - \hat{\mathrm{E}}(X|R)$ is an imprecise estimate of the unmeasured confounder $U$. It remains to be explored whether methods for measurement error correction, such as Simulation-Extrapolation (Carroll et al., 2006), can help attenuate this bias.
In linear structural mean models (Robins, 1994), the assumption that the treatment effect is not modified by the instrument, i.e. that

$$E(Y - Y_0|X, R) = \psi^* X$$

does not depend on $R$, implies that marginal and conditional effects are the same. This is seen because the above model implies the same observed data restriction, namely $E(Y - \psi^* X|R) = E(Y - \psi^* X)$, as model $E(Y - Y_1|X, R) = \phi^*(X - 1)$, thus indicating that

$$E(Y_1 - Y_0|X = 1, R) \equiv \psi^* = \phi^* \equiv E(Y_1 - Y_0|X = 0, R)$$

and, consequently, that $\psi^* = E(Y_1 - Y_0)$. The same is no longer true in logistic structural mean models, where additional assumptions are required to infer marginal causal effects. Adjustment for baseline covariates (more generally, covariates which are not causally affected by exposure, outcome or IV) can be discussed in future work. Covariate adjustment is easily realized for all considered IV-estimators by additionally including these in all considered regression models. For the exact IV-estimators, this requires testing whether the predicted counterfactual outcome $Y_0$ is independent of the IV, conditional on baseline covariates. When there is a continuous baseline covariate or multiple discrete covariates, then the model of Vansteelandt and Goetghebeur (2003) is no longer guaranteed to yield a congenial parameterization, unlike the model of Robins and Rotnitzky (2004). The similarity of the estimating functions indicates that, nevertheless, similar estimates would typically be obtained with both approaches.
Because of the close link between IV-estimators for inferring causal effects and IV-estimators for correcting for measurement error (Dunn, 2005; Greenland, 2000, 2005), it will be of interest to examine in future work whether the considered exact IV-estimators of the causal odds ratio may shed new light on measurement error correction in logistic regression models in the presence of an IV. In Chapter 3, we have taken this link one step further in the context of linear structural mean models, and used one IV for inferring a causal effect and another IV for correcting for measurement error in the exposure. Specifically, we have proposed a general procedure to correct

IV estimators for systematic error in the exposure when an additional instrumental variable for the measurement error was available. Note from the Chapter 3 that the 2 considered IV's satisfy different restrictions. The IV for inferring a causal effect satisfies the assumptions that it is: associated with the exposure, has not direct effect on outcome except through the exposure, and that its association with outcome is unconfounded; and in contrast the IV for measurement error correction satisfies the assumption that it is a (pre-exposure) surrogate for the observed exposure (in the sense that it is correlated with exposure), which was thus assumed not to modify the exposure effect of interest. In our example, the Causal-IV $R$ was randomization and the Measurement error-IV $T$ was age which is assumed not to modify the treatment effect. In placebo-controlled randomized trials with noncompliance, measurements $T$ on run-in placebo compliance may very well meet the assumption of a measurement error IV. With concern for compliance mismeasurement, recording run-in compliance may thus be favorable. More generally, we have shown that IV-Causal's can be used as IV's for the measurement error. On theoretical grounds and on the basis of simulation experiments, we recommend the 'improved error-adjusted estimator' of Chapter 3 for data analysis. This estimator was designed so that adjustment for measurement error does not compromise the power of tests of the causal null. This is attractive, knowing that standard tests of the causal null hypothesis (i.e., that the causal instrument $R$ is independent of outcome) ignore exposure measurements and are thus valid in the presence of measurement error. Because the proposed estimator does not converge uniformly to a normal distribution, we recommend the proposed uniform confidence intervals. For illustrative purposes, we have developed this work under structural mean models which assume linear exposure effects that are not modified by pre-exposure covariates. Extensions to linear structural mean models that allow for effect modification by baseline covariates are methodologically straightforward, but computationally more demanding. Finally, we believe the results of Chapter 3 to be more broadly useful as they suggest, in line with Gustafson (2005), that incorporating a little prior information on a weakly identified nuisance parameter may yield substantial efficiency improvements for the target parameter. Similar ideas may therefore prove useful in related settings (Vansteelandt and Goetghebeur, 2004; Fischer and Goetghebeur, 2004; Ten Have et al., 2007) with weak identification. In addition, our results indicate how such prior information may be adopted in a frequentist analysis.

In this thesis, we have finally investigated the impact of exposure reclassification on the asymptotic bias of different effect estimators when exposure effects are confounded by measured covariates. We have done this by comparing inverse probability of treatment weighted (IPTW) estimators and doubly robust estimators for the exposure effect in linear marginal structural mean models (MSM) with G-estimators,

and propensity score (PS) adjusted estimators. Our interest in this stems from the fact that we anticipated the different estimators to be differently affected by reclassification error because some of them not only use the exposure as a covariate in a regression model, but also rely on estimates of the expected exposure, given background covariates. To allow for measurement error correction, it would be of interest to adapt the misclassification SIMEX (MC-SIMEX) approach to settings where the reclassification probabilities are supposed to be known or can be obtained from other studies.

# References


Abadie A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, **113**, 231–263.

Albert J.M. (2008). Mediation analysis via potential outcomes models. *Statistics in Medicine*, **27**, 1282–1304.

Amemiya T. (1974). The non-linear two-stage least-squares estimator. *Journal of Econometrics*, **2**, 105–110.

Amemiya Y. (1990). Instrumental variable estimation of the nonlinear measurement error model. *Journal of Econometrics*, **28**, 273–289.

Amemiya Y. (1990). Two-stage instrumental variable estimation of the nonlinear errors-in-variables model. *Journal of Econometrics*, **44**, 311–322.

Andrews D.W.K. and Ploberger W. (1994). Optimal tests when a nuisance parameter is present only under alternative. *Econometrica*, **62**, 1383–1414.

Angrist J.D., Imbens G.W. and Rubin D.B. (1996). Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association*, **91**, 444–455.

Bashir S.A., Duffy S.W. and Qizilbash N. (1997). Repeat measurements of case-control data: correction for measurement error in a study of ischameic stroke and haemostatic factors. *International Journal of Epidemiology*, **26**, 64–70.

Bickel P.J., Klaassen C.A.J., Ritov Y. and Wellner J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag, New-York.

Brookhart M.A., Wang P.S., Solomon D.H. and Schneeweiss S. (2006). Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology*, **17**, 268–275.

Brookhart M.A. and Schneeweiss S. (2007). Preference-based instrumental variable methods for the estimation of treatment effects: Assessing validity and interpreting results. *The International Journal of Biostatistics*, **3**, Article 14.

Brumback B., Greenland S., Redman M., Kiviat N. and Diehr P. (2003). The intensity-score approach to adjusting for confounding. *Biometrics*, **59**, 274–285.

Budtz-Jørgensen E., Keiding N., Grandjean P., Weilhe P. and White R.F. (2003). Consequences of exposure measurement error for confounder identification in environmental epidemiology. *Statistics in Medicine*, **22**, 3089–3100.

Buonaccorsi J.P., Laake P. and Veierod M.B. (2005). On the effect of misclassification on bias of perfectly measured covariates in regression. *Biometrics*, **61**, 831–836.

Buzas J.S. (1997). Instrumental variable estimation in nonlinear measurement error models. *Communications in Statistics-Theory and Methods*, **26**, 2861–2877.

Buzas J.S. and Stefanski L.A. (1996). Instrumental variable estimation in generalized linear measurement error models. *Journal of the American Statistical Association*, **91**, 47–62.

Buzas J.S., Stefanski L.A. and Tostesson T.D. (2005). Measurement error. www.springerlink.com.

Campbell A. (2002). The potential role of aluminium in Alzheimer's disease. *Nephrol Dial Transplant*, **17**, 17–20.

Carroll R.J. (1989). Covariance analysis in generalized linear measurement errors. *Statistics in Medicine*, **8**, 1075–1093.

Carroll R.J., Küchenhoff H., Lombard F. and Stefanski L.A. (1996). Asymptotics for the SIMEX estimator in nonlinear measurement error models. *Journal of the American Statistical Association*, **91**, 242–250.

Carroll R.J., Ruppert D., Crainiceanu C.M., Tosteson T.D. and Karagas M.R. (2004). Nonlinear and nonparametric regression and instrumental variables. *Journal of the American Statistical Association*, **99**, 736–750.

Carroll R.J., Ruppert D., Stefanski L.A. and Crainiceanu C.M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, Second Edition. CRC Press.

Carroll R.J. and Stefanski L.A. (1990). Approximate Quasi-likelihood estimation in models with surrogate predictor. *Journal of the American Statistical Association*, **85**, 652–663.

Casas J.P., Bautista L.E., Smeeth L., Sharma P. and Hingorani A.D. (2005). Homocysteine and stroke: evidence on a causal link from Mendelian randomization. *Lancet*, **365**, 224–232.

Chu H., Wang Z., Cole S.R. and Greenland S. (2006). Sensitivity analysis of misclassification: A graphical and a bayesian approach. *Annals of Epidemiology*, **16**, 834–841.

Cole S.R. and Chu H. (2005). Effect of acyclovir on herpetic ocular recurrence using a structural nested model. *Contemporary Clinical Trials*, **26**, 300-310.

Cook J.R. and Stefanski L.A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, **89**, 1314–1328.

Curtis K.M., Mohllajee A.P., Martins S.L. and Peterson H.B. (2006). Combined oral contraceptive use among women with hypertension: a systematic review. *Contraception*, **73**, 179-188.

Davies R.B. (1977). Hypothesis testing when a nuisance parameter is present only under alternative. *Biometrika*, **64**, 247–254.

Davies R.B. (1987). Hypothesis-testing when a nuisance parameter is present only under the alternative. *Biometrika*, **74**, 33–43.

Dawson R. and Lavori P.W. (2002). Using inverse weighting and predictive inference to estimate the effects of time-varying treatments on the discrete-time hazard. *Statistics in Medicine*, **21**, 1641–1661.

Didelez V. and Sheehan N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, **16**, 309–330.

Didelez V., Meng S. and Sheehan N. (2008). On the bias of IV estimators for Mendelian randomisation. Technical Report. www.maths.bris.ac.uk/research.

Divers J., Vaughan L.K., Padilla M.A., Fernandez J.R., Allison D.B., et al. (2007). Correcting for measurement error in individual ancestry estimates in structured association tests. *Genetics*, **176**, 1823–1833.

Dunn G. (1999). The problem of measurement error in modeling the effect of compliance in a randomized trial. *Statistics in Medicine*, **18**, 2863–2877.

Dunn G. (2005). *Instrumental Variables.* In Encyclopaedic Companion to Medical Statistics (eds. Everitt, B. and Palmer, C. R.), Arnold Publishers.

Dunn G. (2007). Regression models for methods comparison data. *Journal of Biopharmaceutical Statistics*, **17**, 739-756.

Dunn G. and Bentall R. (2007). Modelling treatment-effect heterogeneity in randomized controlled trials of complex interventions (psychological treatments). *Statistics in Medicine*, **26**, 4719-4745.

Dunn G., Maracy M. and Tomenson B. (2005). Estimating treatment effects from randomized clinical trials with noncompliance and loss to follow-up: the role of instrumental variable methods. *Statistical Methods in Medical Research*, **14**, 369–395.

Dunn G. and Roberts C. (1999). Modelling method comparison data. *Statistical Methods in Medical Research*, **8**, 161–179.

Efron B. and Feldman D. (1991). Compliance as an explanatory variable in clinical trials. *Journal of the American Statistical Association*, **86**, 9–17.

Fischer-Lapp K. and Goetghebeur E. (1999). Practical properties of some structural mean analyses of the effect of compliance in randomized trials. *Controlled Clinical Trials*, **20**, 531–546.

Fischer K. and Goetghebeur E. (2004). Structural mean effects of noncompliance: Estimating interaction with baseline prognosis and selection effects. *Journal of the American Statistical Association*, **99**, 918–928.

Fosgate G. (2006). Non-differential measurement error does not always bias diagnostic likelihood ratios towards the null. *Emerging Themes in Epidemiology*, **37**, 1–7.

Foster E.M. (1997). Instrumental variables for logistic regression: an illustration. *Social Science Research*, **26**, 487–504.

Frangakis C.E. and Rubin D.B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment noncompliance and subsequent missing outcomes. *Biometrika*, **86**, 365-79.

Freedman L.S., Midthune D., Carroll R.J. and Kipnis V. (2008). A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression. *Statistics in Medicine*, **27**, 5195–5216.

Fuller W.A. (1987). *Measurement Error Models* . New York: John Wiley.

Glymour M.M., Weuve J. and Chen J.T. (2008). Methodological challenges in causal research on racial and ethnic patterns of cognitive trajectories: measurement, selection, and bias. *Neuropsychology Review*, **18**, 194–213.

Goetghebeur E. and Lapp K. (1997). The effect of treatment compliance in a placebo-controlled trial: regression with unpaired data. *Journal of the Royal Statistical Society*, **46**, 351–364.

Goetghebeur E. and Shapiro H.S. (1996). Analyzing non-compliance in clinical trials: ethical imperative or mission impossible?. *Statistics in Medicine*, **15**, 2813–2826.

Goetghebeur E., Molenberghs G. and Katz J. (1998). Estimating the causal effects of compliance on binary outcome in randomized controlled trials. *Statistics in Medicine*, **17**, 341–355.

Goetghebeur E. and Vansteelandt S. (2005). Structural mean models for compliance analysis in randomized clinical trials and the impact of errors on measures of exposure. *Statistical Methods in Medical Research*, **14**, 397–415.

Greene W.H. (2003). *Econometric Analysis (5th edn)*. Prentice-Hall: Englewood Cliffs.

Greenland S. (2000). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, **29**, 722–729.

Greenland S. (2005). Multiple-bias modeling for analysis of observational data. *Journal of the Royal Statistical Society*, **168**, 267–291.

Greenland S. and Gustafson P. (2006). Accounting for independent nondifferential misclassification does not increase certainty that an observed association is in the correct direction. *American Journal of Epidemiology*, **164**, 63–68.

Greenland S., Pearl J. and Robins J.M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, **10**, 37–48.

Gustafson P. (2003). *Measurement Error and Misclassification in Statistics and Epidemiology Impacts and Bayesian Adjustments*. Press/CRC.

Gustafson P. (2005). On model expansion, model contraction, identifiability and prior information: two illustrative scenarios involving mismeasured variables. *Statistical Science*, **20**, 111–129.

Gustafson P. and Le N.D. (2002). Comparing the effects of continuous and discrete covariate mismeasurement, with emphasis on the dichotomization of mismeasured predictors. *Biometrics*, **58**, 878–887.

Gustafson P., Le N.D. and Saskin R. (2001). Case-control analysis with partial knowledge of exposure misclassification probabilities. *Biometrics*, **57**, 598–609.

Guolo A. (2008). Robust techniques for measurement error correction: a review. *Statistical Methods in Medical Research*, **17**, 555–580.

Hansen B.E. (1992). The likelihood ratio test under nonstandard conditions - testing the Markov switching model of GNP. *Journal of Applied Econometrics*, **7**, S61–82.

Hausman J.A., Neweya W.K. and Powell J.L. (1995). Nonlinear errors in variables estimation of some Engel curves. *Journal of Econometrics*, **65**, 205–233.

Heckman J.J. and Robb R.J. (1985). Alternative methods for evaluating the impact of interventions : An overview. *Journal of Econometrics*, **30**, 239–267.

Heid I.M., Küchenhoff H., Wellmann J., Gerken M., Kreienbrock L., et al. (2002). On the potential of measurement error to induce differential bias on odds ratio estimates: an example from radon epidemiology. *Statistics in Medicine*, **21**, 3261–3278.

Henneman T.A., van der Laan M.J. and Hubbard A.E. (2002). Estimating causal parameters in marginal structural models with unmeasured confounders using instrumental variables. (U.C. Berkeley Division of BiostatisticsWorking Paper Series, paper 104). Berkeley, CA: The Berkeley Electronic Press.

Hernán M.A. (2004). A definition of causal effect for epidemiological research. *Journal for epidemiology and community health*, **58**, 265–271.

Hernán M.A., Brumback B.A. and Robins J.M. (2002). Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. *Statistics in Medicine*, **21**, 1689–1709.

Hernán M.A. and Robins J.M. (2006). Instruments for causal inference-An epidemiologist's dream? *Epidemiology*, **17**, 360–372.

Höfler M. (2005). The effect of misclassification on the estimation of association: a review. *International Journal of Methods in Psychiatric Research*, **14**, 92–101.

Holland P.W. (1986). Statistics and Causal Inference. *American Sociological Association*, **81**, 945–960.

Holland P.W. (1988). *Causal Inference, Path Analysis, and Recursive Structural Equations Models (with Discussion)*. In: Clogg CC ed. Sociological methodology 1988. Washington DC: *American Sociological Association*, pp. 449-484.

Hosmer D.W. and Lemeshow S. (2000). *Applied Logistic Regression*. New York Wiley, Series in probability and Statistics.

Hui S.L. and Walter S.D. (1980). Estimating the error rates of diagnostic tests. *Biometrics*, **36**, 167–171.

Joffe M.M., Small D. and Hsu C.Y. (2007). Defining and estimating intervention effects for groups that will develop an auxiliary outcome. *Statistical Science*, **22**, 74–97.

Johnston K. M., Gustafson P., Levy A.R. and Grootendorst P. (2008). Use of instrumental variables in the analysis of generalized linear models in the presence of unmeasured confounding with applications to epidemiological research. *Statistics in Medicine*, **27**, 1539–1556.

Jurek A.M., Greenland S., Maldonado1 G. and Church T.R. (2005). Proper interpretation of non-differential misclassification effects: expectations vs observations. *International Journal of Epidemiology*, **34**, 680-687.

Khaw K.T., Day N., Bingham S., Wareham N. (2004). Observational versus randomised trial evidence. *Lancet*, **364**, 753–4.

Kipnis V., Subar A.F., Midthune D., Freedman L.S., Ballard-Barbash R., et al. (2003). The structure of dietary measurement error: results of the OPEN biomarker study. *American Journal of Epidemiology*, **158**, 14–21.

Koul H.l. and Song W. (2008). Regression model checking with Berkson measurement errors. *Journal of Statistical Planning and Inference*, **138**, 1615–1628.

Küchenhoff H., Lederer W. and Lesaffre E. (2007). Asymptotic variance estimation for the misclassification SIMEX. *Computational Statistics and Data Analysis*, **51**, 6197–6211.

Landin R., Freedman L.S. and Carroll R.J. (1995). Adjusting for time trends when estimating the relationship between dietary intake obtained from a food frequency questionnaire and true average intake. *Biometrics*, **51**, 169–181.

Lawlor D.A., Harbord R.M., Sterne J.A.C., Timpson N. and Smith G.D. (2008). Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine* **27**, 1133–1163.

Lederer W. and Küchenhoff H. (2006). A short Introduction to the SIMEX and MCSIMEX. Technical Report. www.stat.uni-muenchen.de.

Leigh J.P. and Schembri M. (2004). Instrumental variables technique: cigarette price provided better estimate of effects of smoking on SF-12. *Journal of Clinical Epidemiology*, **57**, 284–293.

Lewbel A. (2007). Estimation of average treatment effects with misclassification. *Econometrica*, **75**, 537–551.

Lewbel A. (2007). Endogenous selection or treatment model estimation. *Journal of Econometrics*, **141**, 777–806.

Liang K.Y. and Zeger S.L. (1986). Longitudinal data analysis using generalized linear model. *Biometrika*, **73**, 13–22.

Lunceford J.K. and Davidian M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, **23**, 2937–2960.

Lyles R.H., Zhang F. and Drews-Botsch C. (2007). Combining internal and external validation data to correct for exposure misclassification a case study. *Epidemiology*, **8**, 321–328.

MacMahon S., Peto R., Cutler J., Godwin DPhil J., MacMahon S., et al. (1990). Blood pressure, stroke, and CHD. Part 1, prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. *Lancet*, **335**, 765–774.

Mahajan A. (2006). Identification and estimation of regression models with misclassification. *Econometrica*, **74**, 631–665.

Martens E.P., Pestman W.R., de Boer A., Belitser S.V. and Klungel O.H. (2006). Instrumental variables application and limitations. *Epidemiology*, **17**, 260–267.

McCullagh P. and Nelder J.A. (1989). *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.

Minelli C., Thompson J.R., Tobin M.D. and Abrams K.R. (2004). An integrated approach to the meta-analysis of genetic association studies using Mendelian randomization. *American Journal of Epidemiology*, **160**, 445–452.

Moore K.L. and van der Laan M.J. (2007). Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation. U.C. Berkeley Division of Biostatistics Working Paper Series, Paper 215.

Morrissey M.J. and Spiegelman D. (1999). Matrix methods for estimating odds ratios with misclassified exposure data. *Biometrics*, **55**, 338–44.

Nagelkerke N., Fidler V., Bernsen R. and Borgdorff M. (2000). Estimating treatment effects in randomized clinical trials in the presence of non-compliance. *Statistics in Medicine*, **19**, 1849-1864.

Nananda F., Col M.D., Stephen G. and Pauker M.D. (2003). The discrepancy between observational studies and randomized trials of menopausal hormone therapy: did expectations shape experience? *Annals of Internal Medicine*, **139**, 923–929.

Neuhaus J.M. (1999). Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika*, **86**, 843–855.

Neyman J. (1923). *On the Application of Probability Theory to Agricultural Experiments*. Essay on Principles. Section 9, translated in Statistical Science, (with discussion), Vol 5, No 4, 465.480, 1990.

Palmer T.M., Thompson J.R., Tobin M.D., Sheehan N.A. and Burton P.R. (2008). Adjusting for bias and unmeasured confounding in Mendelian randomization studies with binary responses. *International Journal of Epidemiology*, **37**, 1161–1168.

Pearl J. (1995). Causal diagrams for empirical research. *Biometrika*, **82**, 669–710.

Pearl J. (2000). *Causality Models, Reasoning, and Inference*. Cambridge University Press, New York.

Pearl J. (2001). Causal inference in the health sciences: A conceptual introduction. *Health Services and Outcomes Research Methodology*, **2**, 189-220.

Pearl J. (2001). Causal inference in statistics: A gentle introduction. *Computing Science and Statistics, Proceedings of Interface*, **33**, Technical Report.

Pearl J. (2002). A graphical criterion for the identification of causal effects in linear models. *Proceedings of the Eighteenth National Conference on Artificial Intelligence*, 567–573.

Permutt T. and Hebel J.R. (1989). Simultaneous-equation estimation in a clinical trial of the effect of smoking on birth-weight. *Biometrics*, **45**, 619–622.

Prentice R.L., Pettinger M. and Anderson G.L. (2005). Statistical issues arising in the Women's Health Initiative. *Biometrics*, **61**, 899–911.

Rassen J.A., Schneeweiss S., Glynn R.J., Mittleman M.A. and Brookhart M.A. (2009). Instrumental variable analysis for estimation of treatment effects with dichotomous outcomes. *American Journal of Epidemiology*, **169**, 273–284.

Reade-Christopher S.J. and Kupper L.L. (1991). Effects of exposure misclassification on regression analyses of epidemiologic follow-up study data. *Biometrics*, **47**, 535–548.

Reeves G.K., Cox D.R., Darby S.C. and Whitley E. (1998). Some aspects of measurement error in explanatory variables for continuous and binary regression models. *Statistics in Medicine*, **17**, 2157–2177.

Robins J.M. (1986). A new approach to causal inference in mortality studies with sustained exposure period-application to control of the healthy worker survivor effect. *Computers and Mathematics with Applications*, **7**, 1393–1512.

Robins J. M. (1987). A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Diseases*, **40**, 139s–161s.

Robins J. M. (1989). The control of confounding by intermediate variables. *Statistics in Medicine*, **8**, 679–701.

Robins J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics: Theory and Methods*, **23**, 2379–2412.

Robins J. M. (1998). Correction for non-compliance in equivalence trials. *Statistics in Medicine*, **17**, 269–302.

Robins J.M. (1999). Association, causation, and marginal structural models. *Synthese*, **121**, 151–179.

Robins J.M. (1999). Marginal structural models versus structural nested models as tools for causal inference. In Statistical Models in Epidemiology the Environment, and Clinical Trials, Ed. M. E. Halloran and D. Berry, pp. 95–113. New York: Springer.

Robins J.M. (2005). Optimal structural nested models for optimal sequential decisions. In D.Y. Lin and P.J. Haegerty, editors, Proceedings of the second Seattle Symposium in Biostatistics, pages 189–326. Springer Verlag, New York.

Robins J.M., Hernán M.A. and Brumback B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, **11**, 550–560.

Robins J.M., Mark S.D. and Newey W.K. (1992). Estimating exposure effects by modeling the expectation of exposure conditional on confounders. *Biometrics*, **48**, 479–495.

Robins J.M. and Rotnitzky A. (2004). Estimation of treatment effects in randomized trials with non-compliance and a dichotomous outcome using structural mean models. *Biometrika*, **91**, 763–783.

Robins J.M., Rotnitzky A. and Zhao L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, **89**, 846–866.

Robins J.M., VanderWeele T.J. and Richardson T.S. (2006). Comment on: Causal effects in the presence of non compliance: a latent variable interpretation. *Metron International Journal of Statistics*, **64**, 288–298.

Rosenbaum P.R. (2002). *Observational Studies*. Second Edition, Springer Series in Statistics.

Rosenbaum P.R. and Rubin D.B. (1983). The central role of the propensity-score in observational studies for causal effects. *Biometrika*, **70**, 41–55.

Rosner B. and Colditz G. (1996). Nurses' health study: log-incidence mathematical model of breast cancer incidence. *Journal of the National Cancer Institute*, **88**, 359–646.

Rosner B., Spiegelman D. and Willett W.C. (1990). Correction of logistic regression relative risk estimates and confidence intervals for random within-person measurement error. *American Journal of Epidemiology*, **136**, 1400–1413.

Rosner B., Willett W.C. and Spiegelman D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine*, **8**, 1051–1070.

Rubin D.B. (1997). Estimating causal effects from large data sets using propensity-scores. *Annals of internal Medicine*, **127**, 757–763.

Rubin D.B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, **66**, 688–701.

Rubin D.B. (1978). Bayesian-inference for causal effects-role of randomization. *Annals of Statistics*, **6**, 34–58.

Rubin D.B. (1990). Formal modes of statistical inference for causal effect. *Journal of Statistical Planning and Inference*, **25**, 279–292.

Rudemo M., Ruppert D. and Streibig J.C. (1980). Random effect models in nonlinear regression with applications to bioassay. *Biometrics*, **45**, 349–362.

Serfling R.J. (1980). *Approximation Theorems of Mathematical Statistics.* Wiley: New York.

Smith G.D. and Ebrahim S. (2004). Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology*, **33**, 30–42.

Smith G.D., Harbord R., Milton J., Ebrahim S. and Sterne J.A.C. (2005). Does elevated plasma fibrinogen increase the risk of coronary heart disease? Evidence from a meta-analysis of genetic association studies. *Arteriosclerosis Thrombosis and Vascular Biology*, **25**, 2228–2233.

Sommer A. and Zeger S.L. (1991). On estimating efficiency from clinical trials. *Statistics in Medicine*, **10**, 45–52.

Spiegelman D., Carroll R.J. and Kipnis V. (2001). Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument. *Statistics in Medicine*, **20**, 39–160.

Spiegelman D., Rosner B. and Logan R. (2000). Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation study designs. *Journal of the American Statistical Association*, **95**, 51–61.

Spirtes P. Glymour C. and Scheines R. (2000). *Causation, Prediction, and Search*. 2nd edition, MIT Press.

Stefanski L.A. (1985). The effects of measurement error on parameter estimation. *Biometrika*, **72**, 583–592.

Stefanski L.A. and Buzas J.S. (1995). Instrumental variables estimation in binary regression measurement error models. *Journal of The American Statistical Association*, **90**, 541–550.

Stefanski L.A. and Carroll R.J. (1991). Deconvolution-based score tests in measurement error models. *Annals of Statistics*, **19**, 249–259.

Ten Have T.R., Elliott M.R., Joffe M., Zanutto E. and Datto C. (2004). Causal models for randomized physician encouragement trials in treating primary care depression. *Journal of The American Statistical Association*, **99**, 16–25.

Ten Have T.R., Joffe M. and Cary M. (2003). Causal logistic models for non-compliance under randomized treatment with univariate binary response. *Statistics in Medicine*, **22**, 1255–1283.

Ten Have T.R., Joffe M.M., Lynch K.G., Brown G.K., Maisto S.A. and Beck A.T. (2007). Causal mediation analyses with rank preserving models. *Biometrics*, **63**, 926-934.

Thomas D.C. and Conti D.V. (2004). Commentary: The concept of Mendelian randomization. *International Journal of Epidemiology*, **33**, 21–25.

Thompson J.R. and Carter R.L. (2007). An overview of normal theory structural measurement error models. *International Statistical Review*, **75**,183–198.

Thompson J.R., Tobin M.D. and Minelli C. (2003). On the accuracy of estimates of the effect of phenotype on disease derived from Mendelian randomization studies. Technical Report.GE1. www.prw.le.ac.uk. Leicester: University of Leicester

Thürigen D., Spiegelman D., Belttner M., Heuer C. and Bernner H. (2000). Measurement error correction using validation data: a review of methods and their applicability in case-control studies. *Statistical Methods in Medical Research*, **9**, 447–474.

Urquhart J. and De Klerk E. (1998). Contending paradigms for the interpretation of data on patient compliance with therapeutic drug regimens. *Statistics in Medicine*, **17**, 251–267.

Van Damme L., Ramjee G., Alary M., Vuylsteke B., Chandeying V., et al. (2002). Effectiveness of COL-1492, a nonoxynol-9 vaginal gel, on HIV-1 transmission in female sex workers: a randomized controlled trial. *Lancet*, **360**, 971–977.

van der Vaart A.W. (1998). *Asymptotic Statistics*. Cambridge University Press: Cambridge.

van der Laan M.J. and Robins J.M. (2002). *Unified Methods for Censored Longitudinal Data and Causality*. Series, Springer Series in Statistics.

VanderWeele T.J., Hernán A.M. and Robins J.M. (2008). Causal directed acyclic graphs and the direction of unmeasured confounding bias. *Epidemiology*, **19**, 720-728.

Vansteelandt S. (2007). On confounding, prediction and efficiency in the analysis of longitudinal and cross-sectional clustered data. *Scandinavian Journal of Statistics*, **34**, 478–498.

Vansteelandt S., Babanezhad M. and Goetghebeur E. (2007). Correcting instrumental variables estimators for systematic measurement error. Harvard University Biostatistics Working Paper Series, Paper 70.

Vansteelandt S. and Goetghebeur E. (2004). Using potential outcomes as predictors of treatment activity via strong structural mean models. *Statistica Sinica*, **14**, 907–925.

Vansteelandt S. and Goetghebeur E. (2003). Causal inference with generalized structural mean models. *Journal of the Royal Statistical Society, Series B*, **65**, 817–835.

Vansteelandt S. and Goetghebeur E. (2005). Sense and sensitivity when correcting for observed exposures in randomized clinical trials. *Statistics in Medicine*, **24**, 191–210.

Vansteelandt S., Goetghebeur E., Kenward M. G. and Molenberghs G. (2006). Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statistica Sinica*, **16**, 953–979.

Vansteelandt S., Rotnitzky A. and Robins J.M. (2007). Estimation of regression models for the mean of repeated outcomes under non-ignorable non-monotone non-response. *Biometrika*, **94**, 841–860.

Veierod M.B. and Laake P. (2001). Exposure misclassification: bias in category specific poisson regression coefficients. *Statistics in Medicine*, **20**, 771–784.

Wang L. (2004). Estimation of nonlinear models with Berkson measurement errors. *Annals of Statistics*, **32**, 2559–2579.

White I.R. (2005). Uses and limitations of randomization-based efficacy estimators. *Statistical Methods in Medical Research*, **14**, 327–347.

Wooldridge J.M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA:MIT Press.

Wooldridge J.M. (2003). *Introductory Econometrics: A Modern Approach* (2nd edition). Cinicinatti, OH. South-Western College Publishing.

Yu Z. and van der Laan M.J. (2003). Double robust estimation in longitudinal marginal structural models. *the Journal of Statistical Planning and Inference*, **136**, 1061–1089.